# ESSAYS ON FINANCIAL ECONOMICS

A Dissertation
Presented to
The Academic Faculty

by

He Zhang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Economics in the
School of Economics

Georgia Institute of Technology
May 2020

# ESSAYS ON FINANCIAL ECONOMICS

Approved by:

Dr. Patrick McCarthy, Advisor
School of Economics
*Georgia Institute of Technology*

Dr. Willie Belton
School of Economics
*Georgia Institute of Technology*

Dr. Haizheng Li
School of Economics
*Georgia Institute of Technology*

Dr. Narayanan Jayaraman
Scheller College of Business
*Georgia Institute of Technology*

Dr. Usha Nair-Reichert
School of Economics
*Georgia Institute of Technology*

Date Approved: Apr 20, 2020

# ACKNOWLEDGEMENTS

I want to extend my gratefulness to my advisor Dr. Patrick McCarthy because this work wouldn't even be possible without him. Dr. Patrick McCarthy has been more than a thesis advisor to me: He is also a life mentor. A special thanks to Dr. Haizheng Li for his support, guidance, and advice as well.

Besides, I want to say thank you to all other members of my thesis committee, Dr. Usha Nair-Reichert, Dr. Willie Belton, and Dr. Narayanan Jayaraman, for the time and efforts. It was my fortune to have them as significant parts of this journey.

I also want to extend my gratefulness to my colleagues, Tony Hardings, Alaina Totten, and Tongyang Yang, for their help and support when I was feeling helpless and isolated. Last but not least, I want to say thank you to my spouse for everything.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS AND ABBREVIATIONS

| | |
|---|---|
| 2SLS | Two-stage least square estimation |
| BLP | Berry- Levinsohn - Pakes demand estimation algorithm |
| CAR | Cumulative abnormal return |
| EBIT | Earnings before interest and tax payments |
| FDIC | Federal Deposit Insurance Corporation |
| FE | Fixed-effectss estimation |
| FED | Federal reserve board |
| GAAP | Generally Accepted Accounting Principles |
| GLS | General least square estimation |
| GMM | Generalized method of moments |
| IFRS | International Financial Reporting Standards |
| IIA | Independent and irrelevant alternatives |
| IPO | Initial public offering |
| IT | Information technology |
| IV | Instrumental variable |
| LASSO | Least absolute shrinkage and selection operator |
| MARS | Multiple adaptive regression spline |
| MBV | Market-to-book value ratio |
| MNL | Multinomial logit model |
| NYSE | New York Stock Exchange |
| OLS | Ordinary least square estimation |
| R&D | Research and development expenses |

| | |
|---|---|
| RE | Random-effects estimation |
| ROA | Return on assets |
| ROE | Return on equities |
| ROIC | Return on invested capital |
| RSS | Residual sum of squared |

# SUMMARY

This dissertation explores the problem of what kind of impacts can education of employees have on firms' financial and economical performances, the problem of how monetary policy made by Federal Reserve Bank can influence the market share of banks in the loan market, and the problem of why Chinese firms give up the premium and IPO overseas.

Building upon the previous research, this dissertation proposes the methodology of combining machine learning and econometrics to answer the questions that both finance researchers and economists don't have answers to. The methodological innovation of this dissertation includes the non-linear control function approach, the inference for marginal effects of cubic polynomial variables, and methods of text mining on financial reports.

By using these methods/approaches proposed, this dissertation finds out that the education of employees have non-linear effects on firms' financial/economic performance, that the interest rate adjustments made by Federal Reserve Bank can squeeze small banks out of the loan market, and that the Chinese firms which IPO overseas are giving up the premium for the government protection and information disclosure requirements. These findings fill in the blanks in the financial economics literature of corresponding areas.

# CHAPTER 1. DO THE COLLEGE DEGREES OF EMPLOYEES HAVE IMPACTS ON FIRM PERFORMANCE: INSIGHTS FROM A NON-LINEAR APPROACH

## 1.1 Introduction

Whether hiring people with higher education levels improves the performance of a firm remains a puzzle: Hiring people with higher education levels implies higher wage costs. Smart (1991) shows that the average salaries of groups with different education levels are significantly different. Laborers with college degrees earn significantly higher salaries than high school graduates. On the other hand, employees with higher education levels are expected to have better skillsets and knowledge foundations, which might increase the overall effectiveness and hence, the overall performance of the firm[1]. Doong et al. (2011) use the data of Taiwan to prove that firms that have more employees with college degrees will have higher market values. In reality, recently many firms believe in degrees and believe that employees with a better educational background can create more profits for the firm, especially in the U.S.

[1] A study by Bontis and Fitzenz (2002) found that the consequences of human capital management and they established the relationship between human capital management and economic and business outcomes. Human capital has a direct impact on the intellectual capital assets that will yield effectiveness measured by four metrics; revenue factor, expense factor, income factor and HC ROI, and a direct impact on higher financial results per employee.

### 1.1.1 The case of China

In China, the human capital issue is an on-going area of active research. Fleisher et al. (2010) discuss the role of human capital in the total factor productivity growth of China; Wang et al. (1999) explore the impact of human capital accumulation on economic growth in China.

As a significant part of human capital, higher education in China has been a primary focus. For that reason, the National Higher Education Entrance Examination[2] has been one of the most critical events for decades. Because the universities independently decide whether to accept the students or not solely based on the student's scores and rankings, regardless of family background, family wealth, and political background, this exam is seen as a fair mechanism of selection. Each year, only the top 45%[3] of the examinees can get into four-year colleges and get bachelor's degrees. Because of the existence of such a "fair" mechanism, most Chinese firms prefer applicants who have bachelors' degrees or even graduate degrees when hiring[4]. Currently, many firms in China define minimum education requirements when hiring, and some require at least a bachelors' degree.[5] Also, Chinese firms usually pay higher salaries to those who have higher levels of education[6]. The demand-side pressure incentivizes more young people to take the National Higher Education Entrance Examination and try to get bachelors'

---

[2] Details of National Higher Education Entrance Examination: https://en.wikipedia.org/wiki/National_Higher_Education_Entrance_Examination
[3] Source: Sohu China education channel. Website: http://www.sohu.com/a/82447900_177136
[4] An article showing the placement statistics of different degree levels in China: http://www.chinadaily.com.cn/interface/toutiaonew/53002523/2016-05-10/cd_25191515.html
[5] Information obtained from the news article published by China Daily on 2015/03/13. The full news article can be accessed through http://www.chinadaily.com.cn/dfpd/jl/2015-03/13/content_19803508.htm
[6] An article proving that average salaries are higher for firms with more employees that have graduate degrees: https://hk.saowen.com/a/0b4f67d8929705b47e020891d4c04bb6635b4fe2cebdfa2f4151e937b502ce79

degrees. These condition helped cultivate an educational service industry [7] that now conditions for about 5% of national GDP.[8]

### 1.1.2 Key question and findings

Given the current situation of the labor market and education system, many questions remain unresolved for the employers: For example, do workers with bachelor's degrees create more profit for the employers? Is it better to have a higher overall education level for the firms? In this paper, I use data from Chinese public firms to empirically explore these questions and provide new insights on the impact that a more educated workforce has on firm performance.

Most previous empirical human capital research focuses primarily on the national level[9]. Though some management literature focuses on the firm level, these studies typically restrict their scope by either studying one or two industries or analyzing the management team or board of directors[10]. This paper differs from these studies in that I examine the firm performance effects associated with the overall education level of the firm's workforce, measured by the proportion of bachelor's degree holders. Further, I explore possible alternative channels through which the overall education level of a firm's workforce can impact a firm's performance.

[7] Generally, the educational service sector in China includes curriculum design, test preparation and training, educational device manufacturing and sales, educational facility management service and admission consulting service. A complete list and detailed information can be found at https://baike.baidu.com/item/%E6%95%99%E8%82%B2%E6%9C%8D%E5%8A%A1

[8] Data sources: http://www.chyxx.com/industry/201609/450644.html

[9] As summarized in the literature review section, Nelson and Phelps(1966); Romer, (1990a); Sergio Rebelo(1991) developed and completed the endogenous growth model incorporating the human capital factor. Romer (1990b), Barro (1991), Mankiw et al. (1992) , Levine and Renelt (1992) and Levine and Zervos (1993) confirms that education, as a part of human capital, does contribute to the economic growth nationally.

[10] As summarized in the literature review section, Hitt et al(2001), Shrader and Siegel (2007) show the education level of the managers and board members have positive correlation with firm performance.

The principal empirical methodology of this paper follows the estimation framework from Murray (1989), Miller and Le Breton-Miller (2006), Li, et al. (2008) and Giroud and Holger (2010), in which they build up reduced-form model to estimate the effects of a specific component in firm operation on firm performance. I use a comprehensive data set that covers the public firms in all industries in China. Based on the empirical results, I find that the proportion of employees having bachelors' degrees within a Chinese public firm has significant positive impacts on its performance. However, the impact is non-linear. I also find out that the source of the improvement of firm performance is the increase in operating income. This paper confirms the positive effects of employee education on firm performance and confirms that the impacts vary from one industry to another. This study also finds that Chinese firms with more college degree holders pursue riskier investments and generate higher operational income.

## 1.2 Literature review

### *1.2.1 A brief view of related literature*

#### 1.2.1.1 <u>Human Capital and Endogenous Growth</u>

Many scholars and economists in macroeconomics tried to identify whether workers' human capital has economic impacts on the whole economy. Nelson and Phelps (1966), Romer (1990a), and Rebelo (1991) developed endogenous growth models from a macro perspective to explain how human capital contributes to the total output of a closed economy. These models not only tell why the level of output is higher when a country has more human capital but also why the growth rate is higher. These models explain why human capital is a crucial factor in the economy, finding that the accumulation of human capital contributes to the high level of innovation, and hence more advanced technology. Those technologies increase productivity, and hence the final output level. In addition, these models propose that education and experience, that is, real-world practice are pathways for increasing and human capital stocks [11].

#### 1.2.1.2 Education and Endogenous Growth

Subsequent theoretical and empirical research in endogenous growth models analysed how education, as a part of the human capital, influences output growth. Romer (1990b), Barro (1991), Mankiw et al. (1992), Levine and Renelt (1992), and Levine and Zervos (1993) use the primary- or secondary-school enrolment rate as a proxy of a country's education level and find that, for nearly all states, these rates have significant and positive impacts on the economy's growth rate[12]. In particular, they find that the school enrolment rate has significant positive effects on the growth of the economy.

Building on this finding, a more recent paper (Hanushek and Kimko, 2000) measures the quality of the labor force using comparative tests of mathematics and scientific skills and finds that math and science skill scores are positively correlated with higher economic growth, which also confirms the positive impact that human capital accumulation has on economic growth. At the same time, many micro- econometricians and labor economists (e.g., Krueger and Lindahl, 2000; and Card, 1999) examined this issue from a microeconomic perspective. Recently, Jones (2014) re-examines the relationship among income difference, human capital level, and the aggregated education

[11] Romer, 1990a points out that learning by doing is also a way to accumulate human capital and it has some spill-over effects over the macroeconomy. Detailed model and derivation can be found in the paper.
[12] The measurement metrics of economic growth do vary across this research. However, the mostly commonly used ones are GDP growth and GNP growth rate.

level of laborers. He proposes a new method for aggregating different human capital skill levels named "Generalized Division of Labor" human capital aggregator. Based on Caselli (2005)'s data, his analysis shows that only when workers are a perfect substitute for each other, the wage can reflect the productivity accurately. Moreover, he uses the new aggregation method to show a significant portion of cross-section differences between countries might be coming from human capital variations, and this portion is bigger than what previous research finds under the perfect substitute assumptions.

Consistent with the findings in the macroeconomic literature, using cross-country analysis, these microeconomic studies find that the change in the education level of a country's general population is positively associated with the shift in economic growth. Though at this stage, many researchers step out from the endogenous growth theory and start to empirically test the effect from individuals' human capital level, measured by education or skill level of labors, on economic growth, their scope is still restricted on the level of country due to the lack of micro-level data.

#### 1.2.1.3 Human capital of firm principals and management

After entering the 21st century, the increasing availability of micro-level data and the development of information technologies enable economic and management researchers to study the human capital effects on firm strategic management. In the business literature, this line of research focuses on how the education level of principals[13] influences the firm's human resource management and firm strategy (e.g., Hitt et al., 2001;

[13] Hitt et al (2001), Hillman (2003), and Shrader and Siegel (2007) define principals of a firm as the middle level managers, high level managers, executive level managers or/and board of director members.

Shrader and Siegel, 2007). Hitt et al. (2001) conducted a regression analysis of law firms. The study uses the law school education and practical experience of the vital principle partners as a proxy for human capital, and they use a profit-revenue ratio to measure the firm's performance[14]. The study finds a positive correlation between a law firm's human capital and firm performance. Shrader and Siegel (2007) analyze the impact of management level characteristics on the performance of venture capital firms. Based on Compustat data, the study finds that the educational and functional area experiences of the managers have positive impacts on the firm's sale growth and rate of return on assets (ROA). Although this research provides new insights on the importunate of human capital and education to firm management, the research is quiet on how the overall educational levels of a firm's non-executive employees, including managers or board members, influence its performance.

1.2.1.4 <u>The labor factor labor and firm performance</u>

Until the early 1980s, there existed few studies examining the impact of employee education level on firm performance. Research in management science started exploring the determinants of firm performance during the 1980s. Hansen and Wernerfelt (1989) propose that person-specific factors, which include labor skills, should have significant impacts on the organizational climate, and hence on overall firm productivity. The study uses Fortune 500 firm data in regression analysis and finds that the labor skills and labor structure have significant positive effects on firm performance, as measured by profitability. Stierwald (2009) seeks to identify factors that influence firm performance.

[14] The legal industry is kind of special therefore, they are using the data reported by the American Lawyer. Details can be found in Hitt et al (2001).

She uses an error correction method to analyze the firm performance impacts of firm productivity and the number of employees. In the study, she finds that the estimated firm productivity and number of employees have positive effects on firm performance measured by profit to asset ratio (which is equivalent to the ROA ratio). Addison (2005) finds out that the presence of worker unions and worker representatives are essential to improving firm performance. In this paper, the author confirms the important role of labor on firm performance, finding that a better labor structure with a worker union can improve the productivity of the firm and increase output and firm revenues. Although focused on the firm, the scope of the role of labor in the production process is relatively general.

#### 1.2.1.5 Labor quality and firm performance

Some other researchers go further and are more specific about the role of labor quality in firm performance: Ton (2009), for example, shows that customer service quality is positively correlated with overall firm profitability. Based on longitudinal data of a large retailer, the study uses conformance quality[15] and service quality as independent variables and profitability of the firm as a dependent variable to conduct a regression analysis. The study finds that both the quality of labor and the number of employees are key factors of conformance quality which is positively related to profitability. In related work, some researchers find a positive correlation between a firm's financial performance and service quality. Ittner and Larcker (1998) find a positive relationship between customer satisfaction (measured by the American Customer Satisfaction Index (ASCI)) and market equity value.

[15] *Conformance quality, according to* Ton (2009), refers to the degree to which a good or service meets certain design standards determined by the employer. It's one of the measures of the employee's performance which uses the objective standards set up by the employer to determine how well an employee is doing.

In a study of 200 firms in the U.S., Anderson et al. (2004) also find a positive relationship between customer satisfaction and market value of a firm (measured as Tobin's Q). Notwithstanding this relationship, the study doesn't explore sources of differences in service qualities across firms.

#### 1.2.1.6 Employee education and firm performance

Recently, a few researchers further explore the relationship between firm performance and human capital. The work that is most similar to this study is Doong et al. (2011), which discusses the relationship between market value and human capital using data from Taiwan. The authors assume that education is input to human capital accumulation. Under this assumption, they use the proportion of employees with some higher education (i.e., some college education) to measure the education level of a firm's employees and to reflect, at least the educational part of the firm's human capital level. The study finds a positive correlation between the employee education level and the market value of the firm.

### *1.2.2 Innovation and contribution*

Though many researchers shed light on the relationship between employee quality and firm performance, few researchers explored the specific relationship between firm performance and firm workforce education level. Briefly, our work is distinct from their work on the following aspects:

1, Previous researchers merely look at the relationship between employee education level and firm market value while this paper mainly looks at the relationship between employee education level and general firm performance, which firm market value is a subset of.

2, This paper uses spline regression, polynomial regression, LASSO regression to get more established empirical models and enhance the accuracy of estimation. In addition to that, this paper also proposes a new control function approach to overcome the non-linearity problem[16].

3, Previous researchers don't go much further to examine the mechanism behind the scene while this paper tries to dig into the rationale behind the scene to see which part of the revenue or which part of the expenditure the cross-section variation of performance is coming from.

The findings of this paper shall contribute to the existing corporate finance and labor economics literatures in the sense that this paper fills in the blank in the literature examining the relationship between employee education level and firm performance. In addition to that, this paper demonstrates the possibility of applying machine learning to economic research. Despite the contribution mentioned above, many questions remain unanswered in the end: Why firms have more employees with a bachelor's degree will gain increases in operational incomes? After taking the self-selection of going public into account, will the findings remain robust? To resolve those puzzles, further economics, psychology, or/and sociology research must be conducted, which are out of the scope of this paper.

[16] The Non-linearity issue here refers to the non-linear structure between the dependent variable and the independent variable of interests, and the non-linear relationship between the instrumental variable and the dependent variable. The proofs and the details can be found in section 1.5.

## 1.3 Theoretical framework

### *1.3.1 Structural framework: short-run (restricted) profit maximization*

The following theoretical model for this paper mainly comes from the classical microeconomics theory. The focus of this paper is on the microeconomic side and to investigate the issue by using the scope of firms. Assuming firms are always solving the following profit maximization problem, and more precisely, in the short-run, firms are solving a restricted profit optimization problem since, in the short-run, not all input factors can be adjusted according to Varian (1992). Based on previous research examined in the literature review section, human capital can be modeled as a stock value that is similar to the physical capital of the firm.[17] Considering a typical firm is trying to maximize the short-run profit in each period by choosing the optimal unskilled labor, and the physical capital and skilled labor are quasi-fixed in this context, immediately:[18]

$$\max \pi_t\left(L_{u,t}; L_{s,t}, \mathrm{K_t}, A_t\right) = p_t Q(L_{u,t}; L_{s,t}, \mathrm{K_t}, A_t) - w_{u,t}L_{u,t} \tag{1}$$

where $p_t$ is the price of goods/services, $Q$ is the quantity function of goods/services sold depending on the inputs and technology level $A_t$ which is exogenous, $w_{u,t}$ is the wage of unskilled labor which is determined by the market equilibrium, $L_{u,t}$ denotes the number of unskilled labor that a firm can choose over to maximize the profit. $L_{s,t}$ denotes the number

[17] Krueger and Lindahl (2000), Card (1999) and Jones (2014) discuss the accumulation on the nation level. In their research they find out that the accumulation of human capital has significant impacts on output level just like the physical capital does.

[18] According to Jones (2014), skilled labors are much less substitutable than unskilled labors. Therefore, it means it's relatively much more difficult to adjust the number of skilled labors hired in short run. Thereby, WE assume in short run, firms don't optimize the profit over skilled labors.

of unskilled labor that a firm has at time t $\mathrm{K_t}$ is the physical capital that the firm has at time t.

When the firm solves for the optimal level for $L_{u,t}$, it will become a non-stochastic function of parameters $p_t, A_t, w_{u,t,}$, with restricted levels of $L_{s,t}, \mathrm{K_t}$.

$$L_{u,t}{}^{*} = f\left(p_t, w_{u,t,}\,;\ L_{s,t}, \mathrm{K_t}, A_t\right) \quad (2)$$

By plugging the optimal level into the restricted indirect profit function:

$$\pi_t\left(L_{u,t}; L_{s,t}, \mathrm{K_t}, A_t\right) = p_t Q(L_{u,t}; L_{s,t}, \mathrm{K_t}, A_t) - w_{u,t} L_{u,t}{}^{*} \quad (3)$$

Therefore, we can model $\pi_t{}^{*}$ as a function of $p_t, A_t, w_{u,t,}$, with restricted levels of $L_{s,t}, \mathrm{K_t}$, we can write the indirect restricted profit function as

$$\pi_t^{*} = F(p_t, w_{u,t,}\,;\ L_{s,t}, \mathrm{K_t}, A_t\ ) \quad (4)$$

where the $F$ denotes the indirect profit function of the firm, and it's a function conditional upon $L_{s,t}, \mathrm{K_t}, A_t$, which is either determined by the market equilibrium level in the short run or exogenous. Thereby, the indirect restricted profit function is a function of $L_{s,t}$ of a firm at time t. The restricted profit function is a short-run case of the general profit function. Therefore, it still inherits the property of the general profit function in the sense that it's non-decreasing in $p_t$ and non-increasing in $w_{u,t}$. Besides, it's convex in $w_{u,t}$ as well.

More specifically, in this paper, I am using the ROA, ROE, and Tobin's Q as the proxies of firms' economic profitability, which reflect performance. By definition

$$\mathrm{ROA} = \frac{\pi_t^{*}}{Total\ book\ value\ of\ assets_t} \quad (5)$$

The amount of physical capital leveraged in production can always be defined as a proportion of total asset, therefore, $m = \frac{\mathrm{K_t}}{Total\ book\ value\ of\ assets_t}$. We can have

$$\text{ROA} = \frac{\pi_t^*}{\frac{1}{m} * \mathrm{K_t}} = h(p_t, w_{u,t,}\ ;\ L_{s,t}, \mathrm{K_t}, A_t) \tag{6}$$

Similarly, for ROE, by definition:

$$\text{ROE} = \frac{\pi_t^*}{Total\ book\ value\ of\ equity_t} \tag{7}$$

Since the amount of physical capital leveraged in production can always be defined as a proportion of total asset, therefore, $m = \frac{\mathrm{K_t}}{Total\ book\ value\ of\ assets_t}$. Since $\varphi = financial\ leverage = \frac{Total\ book\ value\ of\ equity_t}{Total\ book\ value\ of\ assets_t}$ and assume financial leverage is exogenous, we have ROE as

$$\text{ROE} = \frac{\pi_t^*}{Total\ book\ value\ of\ assets_t * Financial\ Leverage} = \frac{\pi_t^*}{\frac{1}{m} * \mathrm{K_t} * \varphi} = \text{g}(p_t, w_{u,t,}\ ;\ L_{s,t}, \mathrm{K_t}, A_t) \tag{8}$$

As for Tobin's Q, by definition:

$$\text{Tobin's } \mathrm{Q}_t = \frac{Total\ market\ value\ of\ assets_t}{Total\ book\ value\ of\ assets_t} \tag{9}$$

$$\text{Tobin's } \mathrm{Q}_t = \frac{Total\ market\ value\ of\ equity_t * 1/Financial\ leverage}{Total\ book\ value\ of\ assets_t}$$

$$= \frac{Total\ market\ value\ of\ equity_t * 1/\varphi}{\frac{1}{\sigma} * \mathrm{K_t}} \tag{10}$$

If the total market value is evaluated on the foundation of P/E ratio $\rho$, which is exogenous,[19] then:

$$Total\ market\ value\ of\ equity_t = \rho * \pi_t^* \tag{11}$$

$$\text{Tobin's Q}_t = \frac{\rho * \pi_t^* * 1/\varphi}{\frac{1}{\sigma} * \mathrm{K_t}} = t(p_t, w_{u,t,}\ ;\ L_{s,t}, \mathrm{K_t}, A_t) \tag{12}$$

Thereby, Tobin's Q, ROA, and ROE are functions of the skilled labor of the firm, as derived above. By using the proportion of college employee that has bachelor's' degree as a proxy for the skilled labor, Tobin's Q, ROA and ROE are functions of the proportion of employees that have bachelors' degrees, which motivates the empirical specification in section 1.5.

### *1.3.2 Conceptual framework from the literature*

Hitt et al. (2001) develop a highly complete conceptual framework regarding the relationship between human capital and firm performance. The study argues that intangible resources are more likely than tangible resources to produce a competitive advantage for firms and that intangible firm-specific resources such as knowledge allow firms to add value to incoming factors of production. They also cite Grant (1996), arguing that knowledge is the most critical competitive asset that a firm possesses, and human capital embodies much of an organization's knowledge. Lane & Lubatkin (1998) and Polanyi (1967) propose to classify knowledge as articulable or tacit: Articulable knowledge can be codified and, therefore, written and easily transferred through education. Labor obtains

[19] PE ratio for a specific firm is usually computed by using the average of P/E ratios of all other firms in the same sector or using the average of P/E ratios of public firms with similar characteristics. Thereby, it's exogenous by nature.

articulable knowledge through formal education (and codified firm practices) while tacit (non-codified) knowledge must be gained from real-world practices through articulable knowledge (Liebeskind, 1996). Hitt, et al (2001) also note that higher education and training usually provide a high level of articulable knowledge in the field of specialty. In addition, D'Aveni (1996) believes that the value of professionals' education often holds throughout one's career and is stable.[20] Then, after the professionals finish their education and training, they apply the explicit knowledge derived from their formal education and build firm-specific tacit knowledge through experience. Hitt et al. (2001) also argue that professionals graduating from the highest-ranked programs in their fields can bring the most human capital to firms through intellectual ability, articulable knowledge, social contacts, and prestige.

Hitt et al. (2001) also recognize that although increasing human capital has many positive benefits for the firm, more human capital incurs higher costs. The study cites Bierman and Gely (1995) that firms usually pay employees more than their marginal productivity early in their careers with the expectation of recouping the investment through high productivity as the employee gains tacit knowledge and learns to apply both articulable and tacit knowledge through practice. In other words, firms believe that, at the time of hire, the present value of expected wages paid to the employee will at least be equal to the present value of the expected productivity of the worker in daily operating activities. Therefore, they suggest that there exists a curvilinear relationship between human capital and firm performance since at the early stage of employment, the wage expenses exceed

[20] In this paper they define the professional education as two parts: 1, the law school, medical school, nursing school, engineering school and business school education experience, 2, the training programs in industrial practices.

the productivity of the employees created while later productivity of employees exceed the wage expenses though the present value of the productivity of employees equals to or be larger than the present value of wage expenses.

The Sankey diagram in Figure 1 summarizes the structural relationship between employee education and firm performance. In fact, this Sankey diagram based on the literature is also reflecting the theoretical model. For instance, the technology in the theoretical model can capture the cumulative tacit knowledge in the sense that employees can help to improve the operational or management process by using the tacit knowledge that they obtained from the daily work. The improvement in the operational or management process definitely enhances the productivity of the firm and can be seen as an improvement in technology.

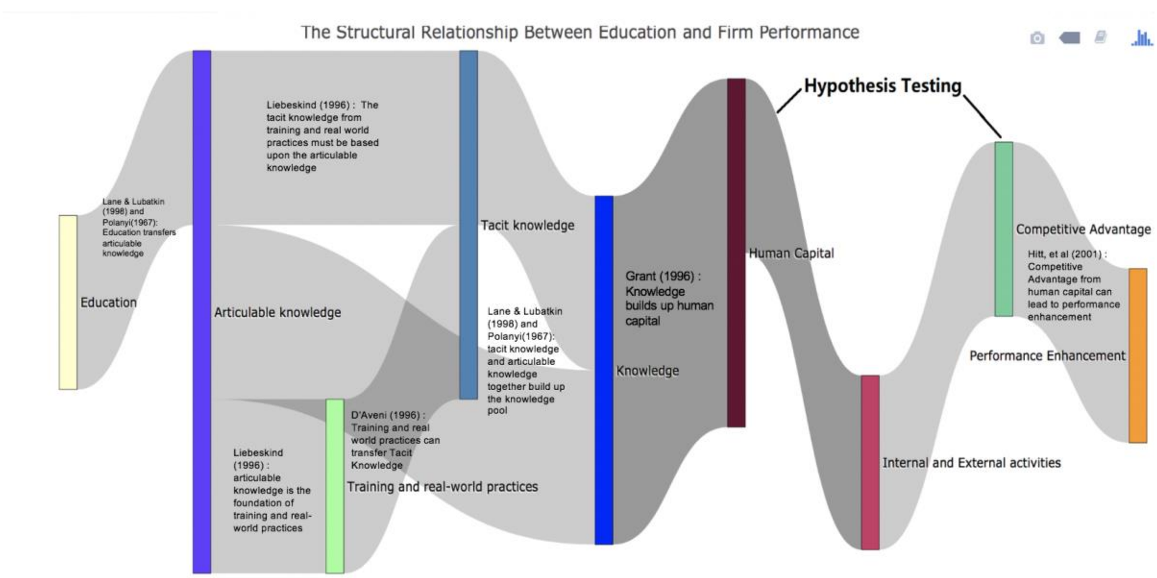


**Figure 1 – A visualization of the structural relationship between education and firm performance.**

*1.3.3 Hypotheses to test*

As demonstrated by the theoretical framework in section 3.1, a typical firm's short-run profit is a function of price level of goods/services and unskilled labor wage level conditional on skilled labor input, physical capital, and technology level. Therefore, the measurement of profitability, namely ROA, is a function of goods/services and unskilled labor wage level conditional on skilled labor input, physical capital, and technology level. The first-order derivative of that function with respect to the skilled labor level equals the marginal effect from the human capital level of the firms on the firm's economic profitability. When proxying the skilled labor level with the proportion of employees with bachelor's degrees in a firm's workforce, we will be able to test whether the skilled labor level, proxied by the proportion of employees with bachelor's degrees, has an impact on the firm's profitability. This leads to the first testable hypothesis:

**H1**: An increase in the proportion of college graduates in a firm's workforce improves a firm's economic profitability, which is measured by the return of asset ratio (ROA).

As indicated by Figure 1 and section 3.2, suggested by the analysis of Bierman and Gely (1995) on the relationship between a firm's human capital and firm performance leads to this study's second hypothesis:

**H2**: The effects of increasing the proportion of employees with bachelor's degrees on firm performance increases at a decreasing rate.

As implied by Figure 1, previous research suggests that the improvements in firm performance should be coming from the competitive advantage generated in its operational activities. Therefore, if the effects of the increase in employee human capital on operational income dominate the increases in operational expenses, then an increase in employee human capital generates competitive advantage for the firm. That gives the third testable hypothesis:

**H3**: Increases in a firm's proportion of employees with bachelor's degrees improve firm performance due to the competitive advantages generated in its operational activities[21].

If increases in employee human capital lead to competitive advantages in operational activities, the competitive advantages likely come from the firm's main business line, rather than auxiliary lines, where the high skill sets obtained from education are transformed into tacit knowledge D'Aveni (1996). That brings the fourth testable hypothesis in this paper:

**H4**: The source of competitive advantages and operational improvements due to increases in a firm's proportion of employees with bachelor's degrees is the firm's primary good/service product line.

As identified by Autor et al. (2007) and Autor et al. (2013), firms have different needs for different input factors since the marginal productivities of the inputs vary across

[21] The competitive advantage in this paper follows the definition given by Hitt et al. (2001). It means the an advantage over competitors gained by offering consumers greater value, either by means of lower prices or by providing greater benefits and service that justifies higher prices.

firms. Therefore, the needs for different types of skills should vary across firms, which lead to the fifth hypothesis to test:

**H5**: The impacts on different firms from the education level of employees are heterogeneous, and increasing the proportion of employees with bachelor's degrees is more likely to have different effects on different firms, especially in the pharmaceutical industries and IT industries.

## 1.4 Data

### *1.4.1 Firm characteristic and employee information data sources*

The firm-level data set in this research is from RESSET database,[22] which many researchers in finance and scholars from Chinese universities recognize as the most comprehensive and reliable source of information on Chinese public companies.

### *1.4.2 Text mining for employee information data*

Even though RESSET contains some information regarding the institutional structure of firm employees, a considerable amount of information is missing.[23] However, many firms provide this information in their annual reports. In order to maximize the sample size and to obtain the maximum amount of information available, I adopted a data mining strategy to collect gain more information on employee education. Appendix B.1 provides details of the data mining strategy.

---

[22] RESSET stands for 锐思数据库 (Ruiswe Database), which can be accessed to through http://www.resset.cn/ with registered/licensed accounts.
[23] Around 30% of the firms in the database don't have the employee education information.

### 1.4.3 Screening and final sample

The sample covers the period from 1999 to 2016. All data are annual. The raw dataset contains 64,583 year-firm combinations. In the data cleaning process, I used several criteria: 1, if there is no data on the overall employee education level- the proportion of employees with bachelor's degrees and the proportion of employees with graduate degrees – then I drop the observation. Since there is no reason to believe that missing annual report data on employee education levels are not random, eliminating these observations should not introduce any bias.[24] Besides, by scrutinizing the dropped data point, there are no clear patterns associated with them: They do not concentrate in a specific industry or location. 2, I drop an observation if the proportion of employees with bachelor's degrees or the proportion of employees with a graduate degree is greater than 1 as this cannot occur. Besides, values lower than 0.0001 are also likely to be miscoded given the prevalence and often a requirement of bachelor's degrees for many managerial positions in public firms. [25] Assuming these mistakes occur randomly, there is no expectation that dropping these will bias the results. After the screening, only 13699 year-firm combinations left.

### 1.4.4 Firm performance measurements

Some researchers provide an empirical framework for research on firm performance. Stierwald (2009) estimates the impacts from a firm-specific factor on performance and provides a guideline of measurements for firm performances. In particular, Stierwald (2009) identifies the following performance metrics: return on assets

[24] See Allison, Paul D. *Missing data*. Vol. 136. Sage publications, 2001.
[25] The management positions defined here include department director, executive level managers, secretaries of the executive level managers, directors on board and general managers. The full details can be found in the news article which can be accessed through http://www.chinadaily.com.cn/dfpd/jl/2015-03/13/content_19803508.htm

(ROA), return on equity (ROE), and earnings before interest payment and tax (EBIT). Safarova, 2010). Villalonga (2004) also proposed that the firm-specific profits (measured by the time-weighted difference between industry ROA and firm ROA) as a measure of firm profitability and, therefore, firm performance. More recently, scholars in management science include market- to- book- value ratio (MBV) and the return on invested capital (ROIC). In general, many researchers (Neely (2002) and Murray (1989), Miller and Le Breton-Miller (2006), Li, et al. (2008) and Giroud and Mueller (2010)) use ROA as a good measurement of profitability as well.

To sum up, the most commonly used firm performance measurements are ROA, MBV ratio, Tobin's Q, ROE, EBIT, and ROIC. Consistent with these researchers, I use ROA as the primary measure of firm performance, and in a series of robustness checks, I also use ROE and Tobin's Q.

Besides, based upon the structural model showing the relationship between short-run profit function and the accounting measures of firm performance,[26] ROA, ROE, and Tobin's Q are functions of skilled labor whose proxy is the proportion of employees in the workforce that have at least bachelor's degrees.

### *1.4.5 Unbalanced panel data*

In order to maximize the sample size, I keep those observations that have missing values on various firm control variables and use these to conduct robustness and sensitivity checks. After data cleaning and screening processes, the sample contains 3,464 firms and

[26] Please see Appendix A.7 for details.

15,480 firm-year observations. This is an unbalanced panel that creates two problems in estimations: First, there is a potential selection problem that reflects the dynamic mechanism of entry and exit. Firms choose whether to go public. To solve this problem, we have to control for the propensity of going public. Due to the absence of the data on non-public firms, estimating the propensity of going public is not possible, which requires alternative methods to handle a potential selection bias. We will have more discussion about this issue as a drawback of this paper in Section 1.9.

### *1.4.6 Summary statistics and non-parametric analysis*

Table 1 reports summary statistics along with a description of the key variables. Panel (a) in Table 1 gives descriptive statistics for the entire sample. Table 1 (b) sums over time and provides summary statistics for the major industries in the example, and panel (c) summarizes over cross-sections and provides summary statistics by year.

## Table 1 – Summary Statistics

Table1(a)[27]

**Summary Statistics**

| Variable | Description | N | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|---|
| ROA | The return on asset ratio. Computed by net income/total asset (%) | 13501 | 6.01 | 5.70 | -3.44 | 59.81 |
| ROE | The return on equity ratio. Computed by net income/total equity (%) | 13481 | 11.40 | 13.94 | -124.35 | 94.80 |
| Tobin's Q | Tobin's Q ratio. Computed by market value of the total asset/book value of total asset. | 11642 | 2.11 | .2.26 | 0.002 | 108.26 |
| bachratio | The proportion of employees that have bachelor's degree. Computed by the number of employee with bachelor's degree/ total employee numbers | 13501 | 0.24 | 0.18 | 0.002 | 0.9749 |
| emp | Number of total employees | 13501 | 5607.54 | 24335.6 | 6 | 548355 |
| AvgSalary | Average wage. Computed by total salary expenses/number of employees (China Yuan) | 13501 | 19866.1 | 74561.5 | 0.1 | 6020369 |
| age | The number of years since establishment. Computed by using the observation year minus the establishing year. | 13501 | 13.95 | 6.17 | 1 | 73 |
| exp | The total expenditure of a specific fiscal year (Million China yuan) | 13501 | 791 | 682 | 1.86 | 27800 |
| markup | The price markup of a specific firm in a specific fiscal year. Computed by net income after depreciation, interest payments divided by total operating revenue. | 13501 | 0.12948 | 0.1217 | 0.00002 | 0.97084 |
| asset | The value of total assets (Million China yuan) | 13501 | 3290 | 4820 | 4.727 | 210000 |
| QTY | The total revenue of other firms in the same industry in the same year. (Million China Yuan) | 13501 | 27200 | 25500 | 2.05 | 72000 |

[27] Notes: 1, The markup stands for the average markup which is computed through total profit divided by total revenue. 2, Data Sources: RESSET database：http://www.resset.cn/

Table 1(b) [28]

**Summary Statistics of firms by Industries**

| | Pharmaceutical | | | IT | | | Others | | |
|---|---|---|---|---|---|---|---|---|---|
| Variable | N | Mean | Std. Dev. | N | Mean | Std. Dev. | N | Mean | Std. Dev. |
| ROA | 7846 | 6.11 | 5.45 | 1193 | 9.74 | 8.42 | 4462 | 4.83 | 4.69 |
| ROE | 7835 | 10.93 | 11.24 | 1193 | 15.60 | 15.66 | 4453 | 11.11 | 17.18 |
| Tobin's Q | 6818 | 2.24 | 2.24 | 742 | 3.74 | 3.10 | 4082 | 1.61 | 1.91 |
| bachratio | 7846 | 0.19 | 0.13 | 1193 | 0.537 | 0.19 | 4462 | 0.26 | 0.18 |
| emp | 7846 | 4406.56 | 9737.72 | 1193 | 2443.91 | 12092.56 | 4462 | 8565.2 | 39653.5 |
| AvgSalary | 7846 | 14830.71 | 79710.93 | 1193 | 14022.68 | 17059.22 | 4462 | 30282.81 | 73555.04 |
| age | 7846 | 13.64 | 6.06 | 1193 | 12.03 | 5.27 | 4462 | 15.07 | 6.36 |
| exp | 7846 | 478 | 1950 | 1193 | 172 | 1110 | 4462 | 1570 | 11700 |
| Markup | 7846 | 0.1095 | 0.097 | 1193 | 0.16 | 0.12 | 4462 | 0.16 | 0.15 |
| asset | 7846 | 6770 | 207 | 1193 | 381 | 2590 | 4462 | 868 | 8350 |
| Q | 7846 | 425000 | 2200000 | 1193 | 19200 | 155000 | 4462 | 69500 | 116000 |

[28] Notes: 1, The markup stands for the average markup which is computed through total profit divided by total revenue. The key assumption for this markup is the firm sell different products in similar quantity, therefore $\sum \pi_j q_j / \sum p_j q_j = \sum \pi_j q / \sum p_j q = \sum \pi_j / \sum p_j = n\bar{\pi} / n\bar{p}$ = $\bar{\pi}/\bar{p}$ = Average Markup.

2, See panel (a) for variable descriptions. The markup is estimated by using the total profit divided by the total revenue.

Table 1(c)

**Summary Statistics by Time Periods**

| | 1999-2004 | | | 2005-2010 | | | 2011-2016 | | |
|---|---|---|---|---|---|---|---|---|---|
| Variable | N | Mean | Std. Dev. | N | Mean | Std. Dev. | N | Mean | Std. Dev. |
| ROA | 1168 | 4.98 | 4.20 | 2912 | 6.09 | 5.55 | 9421 | 6.11 | 5.90 |
| ROE | 1167 | 9.70 | 8.80 | 2905 | 12.68 | 14.56 | 9409 | 11.22 | 11.23 |
| Tobin's Q | 1124 | 2.02 | 1.54 | 2743 | 1.84 | 1.46 | 7775 | 2.22 | 2.55 |
| bachratio | 1168 | 0.19 | 0.17 | 2912 | 0.22 | 0.18 | 9421 | 0.26 | 0.18 |
| emp | 1168 | 3078.81 | 4556.49 | 2912 | 6307.16 | 27672.97 | 9421 | 5704.79 | 24668.8 |
| AvgSalary | 1168 | 21443.39 | 195863.9 | 2912 | 20284.76 | 53238.24 | 9421 | 19541.19 | 48357.03 |
| age | 1168 | 7.91 | 4.51 | 2912 | 11.82 | 4.92 | 9421 | 15.38 | 6.04 |
| exp | 1168 | 191 | 405 | 2912 | 717 | 4830 | 9421 | 919 | 7820 |
| markup | 1168 | 0.11 | 0.11 | 2912 | 0.13 | 0.13 | 9421 | 0.13 | 0.12 |
| asset | 1168 | 411 | 1850 | 2912 | 3240 | 39800 | 9421 | 7820 | 53200 |
| QTY | 1168 | 18300 | 18500 | 2912 | 90600 | 77700 | 9421 | 359000 | 257000 |

As indicated by Table1 (a), the mean proportion of employees that have bachelor degrees is 0.245, indicating that 25% of the employees in Chinese public firms have college degrees. Table 1(b) shows that the information technology (IT) industry has a much higher proportion of employees with bachelors' degrees than the rest of the companies, which is 53%. The pharmaceutical industry has the lowest proportion of employees with bachelors' degrees, which is 19% while this industry takes the most significant portion of the public firms in China.

Table 2 reports the Pearson correlation coefficients between key variables. Table 1(c) implies that the proportion of employees with college degrees is increasing over time. This trend is aligned with the Chinese Educational Ministry's policy of undergraduate enrolment enlargement over the late 90s. Starting in 1998, the Chinese Educational Ministry encouraged colleges and Universities to admit more students. The colleges and universities did double the enrolment from 1999 to 2003, and the enlargement of enrolment continues since 2004. [29] Therefore, there were more and college graduates on the job market during the sample period and firms are taking more and more bachelors' degree holders on the job market. Also, the ROA has increased as well, which suggests a positive correlation between firm performance and the proportion of employees with college degrees.

Figure 2 provides a visualization of the geographical distribution of the companies. As indicated by Figure 2, most of the public firms in the sample are from the south eastern coast of China[30] and the capital city of China. More specifically, more than 80% of the firms in the sample registered their firms in provinces on the south eastern coast or in the capital city of Beijing.

[29] More details can be found in Chen et al. (2004). They elaborate and analyze the transition of the Chinese college admission rate and undergrad enrolments.

[30] The geographical location discussed here refers to the providential location where the firm file its registration. In most cases, this location is the same with the location of the headquarter of the firm. However, in rare cases, they can be different. In china, most of the employees a firm hire will work in their headquarters. Branches are usually leveraged as sales and local operation centers.

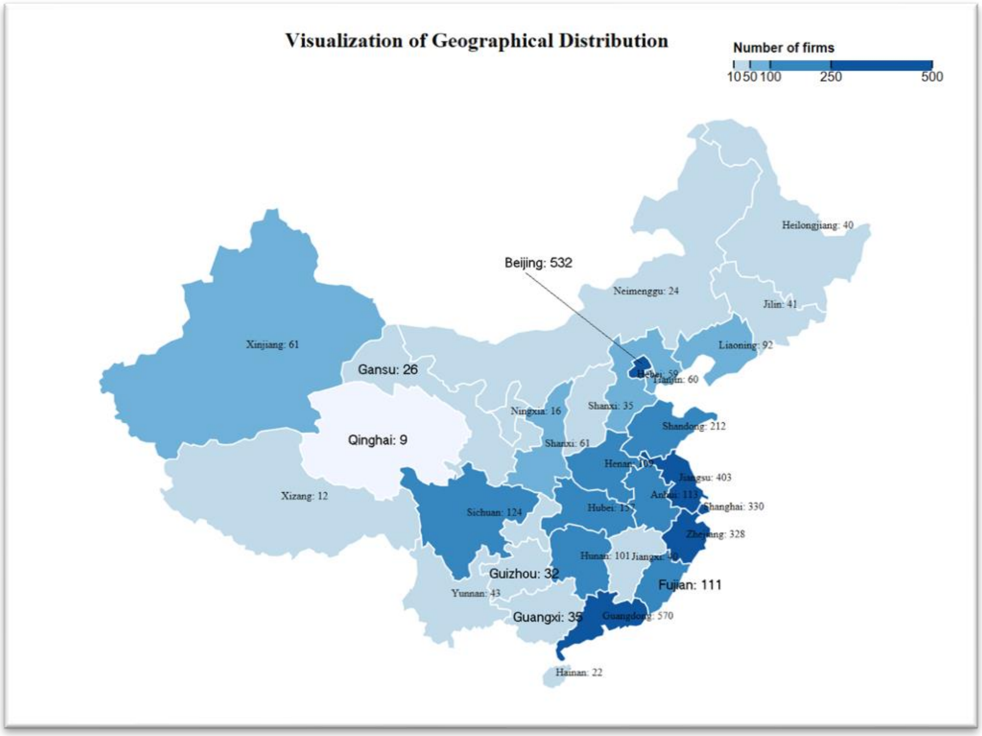


**Figure 2** – **Visualization of the geographical distribution of firms** [31]

Another interesting feature in the sample is the distribution of industry sectors[32]. As shown in Figure 3, the majority of firms in the sample are in the pharmaceutical and Information Technology (IT) industrial sectors. The firms in those two sectors almost make up the whole sample. Some might question the representativeness of the sample due to the industry composition of the sample. However, the distribution of the industry is only telling us that in China, firms in those two industries are more likely to IPO. It does bring up the issue of selection: The industry distribution of private firms in China is significantly different from the distribution of public firms in the sample. For instance, the

[31] An interactive version that contains details of each province can be reached through: https://hzhang440.github.io/ChinaMapCharlesZhang.github.io/

[32] The industry classification is based upon the classification guideline composed by China Securities Regulatory Commission. The complete electronic guideline can be accessed at: http://www.csrc.gov.cn/pub/newsite/scb/ssgshyfljg/201304/W020130402550849843318.doc

pharmaceutical industry only accounts for around 7% of the GDP and accounts for around 10% of the total number of registered firms in China, while in the sample used in this research, the pharmaceutical industry accounts for more than 1/3 of the firms in the sample.[33] Therefore, the conclusion and findings obtained from this dataset should be interpreted with caution; They should not be generalized to all firms in China, and they can be only applied to Chinese public firms.

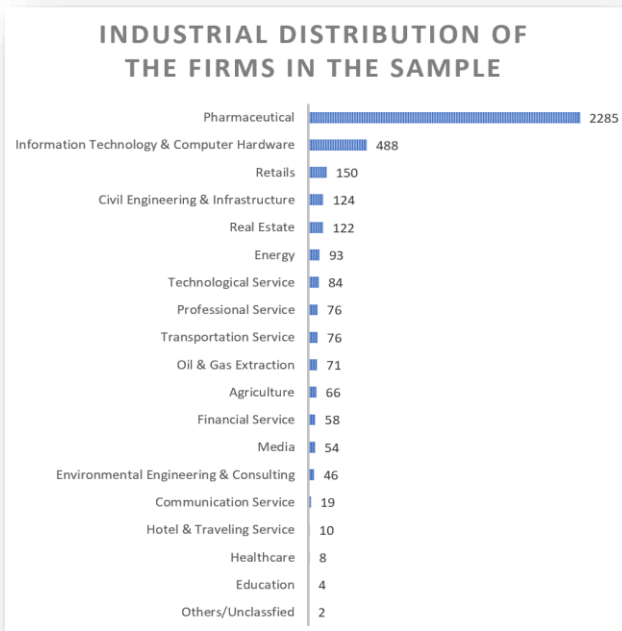


**Figure 3 – Industrial distribution of the firms in the sample**

By looking at the scatter plots and the single-value fitting line in Figure 4, we can't observe a clear linear relationship between the variable ROA and the variable *bachratio* for the data in the years 2007 and 2008. Additionally, taking the Pearson correlation

[33] The complete industry report for the pharmaceutical industry in China can be accessed through https://news.yaozh.com/archive/24765.html.

coefficient into account, though the data in the years 2015 and 2016 shows a seemingly linear pattern with positive slopes, I decide to use non-parametric estimation methods for the relationship between firm performance and its aggregate employee education level since not all the data is following a linear pattern.

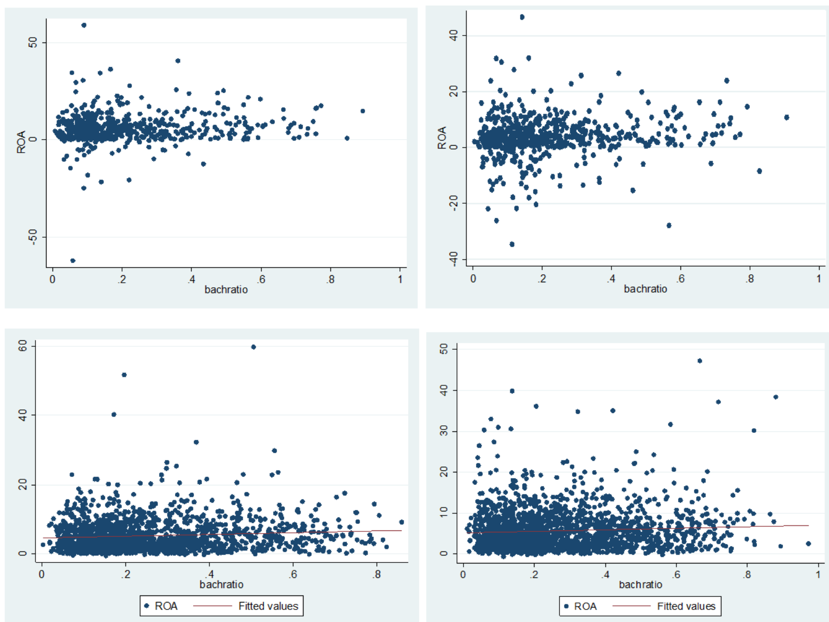


**Figure 4 – Scatter plots and the fitted line between firms' ROA and proportion of employees with bachelors' degrees of year 2007 (upper left), year 2008 (upper right), year 2015 (lower left) and year 2016 (lower right).**

### 1.5 Empirical methodology

This section summarizes the empirical methodology and model specification. Section 1.5.1 discusses a non-parametric functional form for the proportion of a firm's workforce that has bachelor's degrees, f (bachratio). The non-parametric form accounts for potential non-linearities.

Since the model has a large number of variables, there are concerns with potential multicollinearity which suggests a dimension reduction methodology. The technique I adopt is the least absolute shrinkage and selection operator(LASSO) estimation. Section 1.5.2 discusses this technique;

Section 1.5.3 provides a comparison between the linear functional form and the cubic spline functional form;

Although the focus of this study is the impact of a firm's proportion of employees with bachelor's degrees on firm performance, a firm's performance will also affect the number and proportion of college graduates in its workforce. If true and left uncorrected, this endogeneity will bias the estimators. Section 1.5.4 tests for endogeneity and a control function approach is adopted to minimize if not fully account for any endogeneity. The specific empirical specification and procedure for the control function approach are given at the end of section 1.5.4.

The key estimation methods in the analysis are pooled OLS, General Method of Moments (GMM) and fixed-effects estimation. However, the strongly unbalanced panel data for the analysis implies an efficiency problem [34]. For unbalanced panel data, econometric theory[35] indicates that pooled OLS, Fixed-effects (FE) and Random- effects (RE) estimators are unbiased. However, Monte Carlo experiments (Arellano and Bond (1991) and Judson and Owen (1999)) indicate that the efficiency is much lower. Judson

[34] The strongly unbalanced panel data here specifically refers to the panel data with data points that have different length on the time dimension. For instance, some firms in the dataset have only 1 year of length but some firms have all years of length.
[35] Details and proofs can be found in Wooldridge, Jeffrey M. Econometric analysis of cross section and panel data. MIT press, 2010.

and Owen (1999) also find that the most consistent and efficient estimator for short unbalanced panel data is a General Method of Moments (GMM) estimator,[36] which doesn't take the fixed effects of the firm into account. However, those fixed effects might be a source of endogeneity. To ensure the robustness of the results, the baseline results also report the results of a fixed-effects estimation with a control function. This paper doesn't use the random effects (RE) estimator since the RE estimator imposes too many assumptions on the variance-covariance matrix of the error term, and actually, the RE estimator is a special case of General Least Square (GLS) estimator. The goal of GLS is to improve the efficiency of the estimator, which is the same as the goal of the 2-step GMM estimator.

### *1.5.1 Empirical specification*

Based on prior literature and the conceptual framework proposed, the empirical specification leverages the research of Murray (1989), Miller and Le Breton (2006), Li et al. (2008) and Giroud and Mueller (2010). The specification used in estimations is a semi-parametric model

$$y_{it} = a + f(bachratio_{it}) + \boldsymbol{\delta X_{it}} + \boldsymbol{\theta D_{it}} + \varepsilon_{it} \tag{13}$$

In equation (13), $y_{it}$ denotes the dependent variable defined as firm performance, $bachratio_{it}$ is the proportion of employees holding bachelor's degrees. Since the relationship between $y_{it}$ and $educ_{it}$ may not be linear, we adopt a non-parametric form for $educ_{it}$ in this model. $\boldsymbol{X_{it}}$ is the vector of firm control variables, which include the size

[36] Judson and Owen (1999) define short panel data as T=20.

control, number of employees, and firm technology level control-- age. Other control variables based on the theoretical framework in Section 1.3 are:

1, Total Assets to proxy for the physical capital used for production. As indicated in the theoretical model, physical capital used in production can always be modeled as a proportion of the total asset, and the ROA, ROE, Tobin's Q are always functions of physical capital.

2, Average salary controls for the unskilled labor wage. [37]

3, Average markup. Assuming the firms are following monopolist competition industry structures and the consumers have CES preference over the good/services produced by the firms, when expenditure(cost) is controlled and markup is controlled, the aggregate price level of the firm is also controlled[38].

4, Age of firms. The age of firms can be used as a proxy for the technology level of the firm, as indicated by Giroud and Mueller (2010).

5. $D_{it}$ is a vector of fixed effects including Province fixed effects, year fixed effects, and industry fixed effects. Including fixed effects is important given that all determining factors of firm performance are not available: The providential fixed effects together with the year fixed effects are expected to capture consumer demographic and

---

[37] Based on Hamermesh and Grant (1979) and Jones (2014), skilled labor and unskilled labor are substitutable. Therefore, the number of skilled employee can be approximately transformed to a multiple of unskilled labor. That is $N_{skilled} = \vartheta * N_{unskilled}$. The average salary $= \frac{\text{Total salary expenses}}{Total\ amount\ of\ labor} = \frac{\text{Total salary expenses}}{N_{skilled}+N_{unskilled}} = \frac{\text{Total salary expenses}}{\vartheta * N_{unskilled}+N_{unskilled}} = \frac{\text{Total salary expenses}}{(\vartheta+1) * N_{unskilled}} = \frac{\text{Total salary expenses}}{N_{unskilled}} * \frac{1}{(\vartheta+1)} = \frac{1}{(\vartheta+1)}$ *averaege salary of unskilled labor.

[38] Tirole (1988), for example, demonstrates that: the price is a linear function of markup and cost when consumers have CES preference in the monopolistically competitive

policy changes (e.g. production quota) that affect both the firm's performance and the firm's workforce education level. In addition, Provincial fixed effects will capture differences in education infrastructures across provinces. For instance, some provinces have more 985 universities[39] than other provinces, and the firms in those provinces might have easier access to hire college degree holders.

Industry and time fixed effects capture firm-specific production and technology shocks in the sense that if there's a production or technology shock in a specific industry, these effects should have similar impacts on the profitability of all firms in that industry. The industry fixed effect together with the province fixed effect will also capture the difference in accessibility to the labor market of a specific sector.

### *1.5.2 Model selection and dimension reduction*

Among those indicator variables, many of them may be highly correlated since enterprises and firms in a specific geographic area may concentrate in a specific industry due to the comparative advantages that the area offers[40]. Therefore, the Province dummy variables can be highly correlated with a single or a linear combination of the industry dummies. Besides, the year fixed effects are modelled as year dummies in the model. Therefore, a multicollinearity problem may be raised.

---

[39] "985" and "211" universities are the universities that directly report to the Chinese Educational Ministry or the local Providential Educational Ministry. Those universities represent the highest level of higher education in China and they are the most selective universities in general. Therefore, the students who graduate from those universities are seen as the most skillful labors in the market.

[40] For instance, when putting the Beijing location dummy and the IT industry dummy into one regression, the IT industry dummy sometimes get omitted.

In order to resolve that problem, we simply use the model selection technique from unsupervised machine learning: LASSO estimator that Tibshirani (1997) proposed [41]. The intuition of this technique is to add a punishment for including duplicative or unnecessary variables into the estimation specification to prevent overfitting and being too specific to one observation. Then, the criteria function to minimize contains two parts: One is the residual sum of squares (RSS) and the second part is the punishment for including duplicative or unnecessary variables [42]. In order to minimize the criteria function, the estimator eliminates unnecessary or duplicative variables since inducing those variables will add a huge value as a punishment to the criteria function.

We estimate the LASSO estimator on the basic model specification and the results show that most of the location dummies are removed. Briefly, only the dumour variables of the major provinces hosting more than 100 firms remain. The year dummies for only 2000 – 2006, 2008, 2011, 2012, 2014 remain. The LASSO estimator also removes industry dummies for the hotel industry, education industry, and environmental industry.

### *1.5.3 Determine the functional form of education*

As discussed, by looking at the plots using firm performance and aggregated education level as the coordinates, we can't identify a clear linear pattern between these two. Additionally, we compare it by comparing the RSS generated from linear regression and a spline regression. More precisely, we did the following:

[41] Details and rigorous proofs can be found in Tibshiranwe (1997).
[42] In order to minimize the objective function, when the punishment coefficient gets large enough, the optimization solution will be partial to a smaller model, and the coefficients of the unnecessary terms will be suppressed and become 0. In other words, when the coefficient of a term becomes 0, we know it's unnecessary from LASSO.

Since we are exploring the functional form of the relationship between firm performance and aggregated education level, we need to get rid of the effects of other possible factors[43]. More specifically, I did the following:

1, Run an OLS with ROA on all control variables $\boldsymbol{X_{it}}$ and dummy variables after model selection. That is to run OLS on $y_{it} = a + \boldsymbol{\delta X_{it}} + \boldsymbol{\theta D_{it}} + \epsilon_{it}$.

2, We generate the predicted value $\widehat{y_{it}} = \hat{a} + \boldsymbol{\hat{\delta} X_{it}} + \boldsymbol{\hat{\theta} D_{it}}$.

3, After we obtained $\widehat{y_{it}}$, I obtain $y_{it} - \widehat{y_{it}} = \epsilon_{it}$. Where $\epsilon_{it} = f(bachratio_{it}) + \varepsilon_{it}$.

4, We start to compare the fittings of the linear form and non-linear form by doing the following: we run a simple OLS on $\hat{\epsilon}_{it} = \beta_1 bachratio_{it} + \varepsilon_{it}$ and get the RSS.[44]

5, In the meanwhile, we run a spline regression[45] on $\hat{\epsilon}_{it} = f(bachratio_{it}) + \varepsilon_{it}$ and get the RSS. Then we simply compare the RSS of both regressions. The RSS values are plotted in Figure 5:

[43] Proofs can be found in Appendix C, which is based on Bhaumik., et al. (2012)

[44] At this stage, we are focusing on the functional form versus the coefficients. Therefore, the bias problem is not a consideration here.

[45] Details and proofs can be found in De Boor, Carl, et al. A practical guide to splines. Vol. 27. New York: Springer-Verlag, 1978.

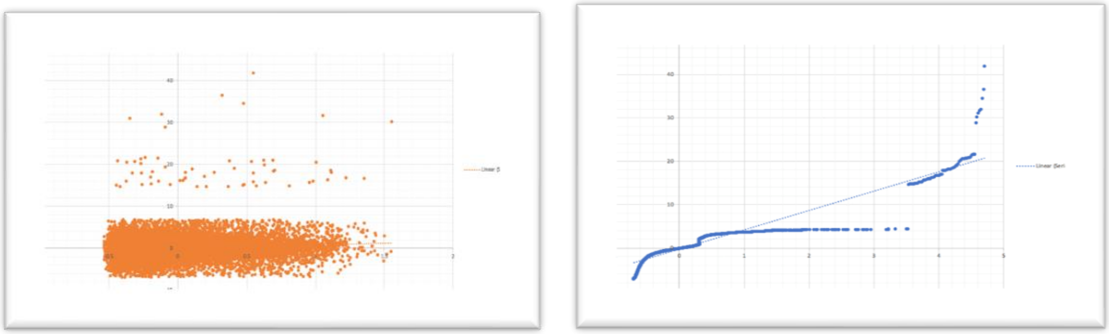

**Figure 5 – Comparison of fitted values of the linear model and cubic smoothing spline.**[46]

Implied by Figure 5, the spline regression provides a much better fit since it gives smaller RSS. Additionally, there are 25 knots in the spline regression and the optimal smoothing punishment is small, which implies that the relationship between firm performance and the proportion of employees with bachelor's degrees is not linear and spline functional form is better for modelling.

However, though a spline regression is good for predictions, it will sacrifice the interpretability and will cause a lot of computational problems when applied to multivariate analysis. Since we know that the smoothing spline function is essentially a cubic polynomial,[47] the $f(bachratio_{it})$ function can be modeled as a third-order polynomial of $bachratio_{it}$. Then specification (1) becomes:

[46] Notes: 1, The left panel contains the fitted value of linear model against the actual value of the "net effect of aggregated education on ROA". The right panel contains the fitted value of cubic smoothing spline model against the actual value of the "net effect of aggregated education on ROA". 2, The RSS of the linear model is 234441.3(N=13319) and the RSS of the cubic smoothing spline is 195323.72344(N=13319).

[47] Proofs of the equivalence between spline and cubic polynomial can be found in De Boor, Carl, et al. A practical guide to splines. Vol. 27. New York: Springer-Verlag, 1978.

$$y_{it} = a + \beta_1 bachratio_{it} + \beta_2 bachratio_{it}{}^2 + \beta_3 bachratio_{it}{}^3 + \boldsymbol{\delta X_{it}} + \boldsymbol{\theta D_{it}} + \varepsilon_{it} \tag{14}$$

### *1.5.4 Identification strategy*

#### 1.5.4.1 Endogenous problem

As previously discussed, a firm's workforce education level is expected to be endogenous. Firms that perform well will able to attract more educated workers with better skill sets because these firms can offer better career development, higher wages, and better benefits.

A Hausman- Durbin- Wu test (Hausman, Durbin and Wu, 1987) to test for endogeneity. The test results confirm that the overall workforce education level measurement, namely the proportion of employees that have bachelor's degrees is endogenous. The test results are shown in Table 2(a).[48]

[48] In the Hausman-Durbin-Wu test, the instrumental variables are borrowed from Hausman and Taylor (1981) and Hausman (2002). WE use the average proportion of employees with bachelor degrees of all other firms in the same provinces (identified by zipcode) and the average proportion of employees with graduate degrees of all other firms the same provinces. By building up a Chi-square statistics between the models using 2SLS and the OLS, the Hausman-Durbin-Wu test will show if the two models are equivalent. If they are not equivalent, it means 2SLS is preferred and endogenous problem exists.

**Table 2 (a) – Hausman-Durbin-Wu test on Endogeneity**

***Hausman-Durbin-Wu test***

| | |
|---|---|
| H0: | IV estimators and OLS estimators are equally consistent |
| H1: | IV estimators and OLS estimators are not equally consistent |
| **chi2(25)** | **= (b-B)'[(V_b-V_B)^(-1)](b-B) = 2646.88** |
| **Prob>chi2** | **0** |

Therefore, if all levels of wages, benefits, incentives as well as other compensations paid to employees, and all levels of skills can be controlled, then the model shouldn't suffer from the endogeneity. However, the data is aggregated on the firm level, and we can't precisely control those factors. Instead, we control the average level of salary to mitigate the endogenous problem. Still, using the reduced form estimation, the coefficient of the education term may be biased. Besides the endogenous tests of the *bachratio* variable, other variables are tested as well. All other variables past the Hausman-Durbin-Wu tests, which means they are not endogenous and won't bias the estimators.

#### 1.5.4.2 Identification strategies

The most common way to resolve the endogenous problem is to use get an instrumental variable (IV). Since we have quite a limited number of variables that can be arguably correlated with the firm's employee educational level and be restricted from the firm performance, we construct a Hausman type IV: The instrumental variables are borrowed from Hausman and Taylor (1981), and Hausman (2002). I use the average proportion of employees with bachelors' degrees of all other firms in the same provinces (identified by province code in the data) in the same provinces[49]. Then, we perform a test regarding the linearity between the IV and the dependent variable by looking at the Pearson correlation coefficient.

**Table 2 (b) – Pearson correlation coefficients**

***Pearson correlation coefficients***
(* denotes significant at confidence level of 0.95)

| | ROA | bachratio | size | emp | salary level | age | expenditure | markup | Q | Hausman -Type IV |
|---|---|---|---|---|---|---|---|---|---|---|
| ROA | 1 | | | | | | | | | |
| bachratio | 0.1703* | 1 | | | | | | | | |
| asset | -0.2776* | -0.0732* | 1 | | | | | | | |
| emp | -0.1867* | -0.2994* | 0.7915* | 1 | | | | | | |
| salary level | 0.0520* | 0.1990* | 0.2902* | 0.0633* | 1 | | | | | |
| age | -0.1658* | -0.0082 | 0.3002* | 0.1767* | 0.1523* | 1 | | | | |
| expenditure | -0.2372* | -0.1661* | 0.8517* | 0.7052* | 0.2252* | 0.2697* | 1 | | | |
| markup | 0.5491* | 0.2249* | -0.1709* | 0.0227* | 0.1759* | -0.017* | -0.2538* | 1 | | |
| Q | -0.0086 | -0.2084* | 0.0879* | 0.1611* | 0.0146 | 0.1895* | 0.1463* | -0.1215* | 1 | |
| Hausman -Type IV | 0.043* | 0.062* | 0.1375* | 0.086* | 0.084* | 0.2297* | 0.1116* | 0.0782* | 0.2461* | 1 |

As shown in Table 2(b), the Pearson correlation coefficient between the dependent variable (ROA) and the Hausman-Type instrumental variable (IV) is quite small in magnitude, which is 0.04, the relationship between these two can't be concluded with a linear relationship. We also perform a test with the linear model against a polynomial model between the two, and the R-sq of the linear model is 0.0018 while the polynomial model's R-sq is 0.0037. Those statistics are telling us that the relationship between the IV and the dependent variable is non-linear[50]. When the instrumental variable (IV) has a non-linear relationship with the dependent variables in a linear model, the instrumental variable

[49] Details can be found in Hausman and Taylor (1981) and Hausman (2002). The rationale of this IV is that other firms' choices might influence this firm's choice of employees but other firms' choices of hiring have nothing to do with this firm's performance.
[50] The Pearson correlation coefficient is 0.04 and the R-sq gets substantially higher if we use 3rd order polynomial regression versus linear regression.

estimator can't be used to identify the parameters[51], which means biases might be introduced if we still use IV estimator or 2SLS procedure to identify the parameters of interests.

Because of that, we follow the control function approach for non-linear models proposed by Powell (2003). More precisely, we use the exogenous control variables along with the Hausman-type IV to predict the fitted value for the endogenous variable and then get the estimator of the error term. By adding the error term into specification (2), I get

$$y_{it} = a + \beta_1 bachratio_{it} + \beta_2 bachratio_{it}{}^2 + \beta_3 bachratio_{it}{}^3 + \omega\hat{v}_{it} + \boldsymbol{\delta X_{it}} + \boldsymbol{\theta D_{it}} + \mathrm{e}_{it} \qquad (15)$$[52]

where $\hat{v}_{it}$ is the estimation of error term when regressing $educ_{it}$ on all other exogenous variables.[53]

Then, a GMM estimation on (15) or an OLS estimation on (3) will give unbiased estimators of all coefficients.[54] After that, the real marginal effect of aggregated employee education level can be identified by looking at $\beta_1$, $\beta_2$ and $\beta_3$, which is $\beta_1 + 2 * \beta_2 * bachratio + 3 * \beta_3{}^2$.

[51] Proofs are attached in Appendix C.
[52] This is under the standard control function approach assumption: $E[e_{it}|educ_{it}] = \omega v_{it}$
[53]There are several factors that differentiate specification (15) from the models used in previous literature:1, Most previous literature presume the independent variables and the dependent variables have linear relationship, which in most cases is not valid;2, Most previous literature either don't pay attention to endogenous problem or simply use instrumental variables to solve the endogenous problem while this model adopts the control function approach. As shown in the proof attached in Appendix C, when the relationship between the instrumental variable and the dependent variable is not linear, IV estimator is no longer valid for identification while control function approach is not limited to this constrain. 3, Previous literature don't take model size into consideration and add all dummy variables in to the model without selection. The dummy variables in speciation (3) are selected through LASSO.
[54] Proofs and details can be found in the control function approach section in Powell (2002).

As shown in Table 2(a), since the correlations between the endogenous variable and the exogenous variables are weak[55]. we can't conclude a linear relationship between them. As mentioned in Woodridge's NBER notes about the control function approach, a linear conditional expectation for the endogenous regressor is a substantive restriction on the conditional distribution of the endogenous regressor. Therefore, when non-linearity exists, the linear projection on endogenous regressor might give wrong estimates of the first stage residual and then might create biases[56]. Therefore, our first stage estimation follows the non-parametric method proposed by Blundell and Powell (2003)[57]. Unlike what Blundell and Powell (2003) and Woodridge's NBER notes suggest, we don't use the kernel regression method because it's difficult to determine the best type of kernel distribution as well as the parameters of bandwidth. Therefore, we use the Multivariate Adaptive Regression Spline (MARS) method,[58] a non-parametric method, in the first stage to obtain $\hat{v}$[59].

More precisely, we did the following:

1, Regress the proportion of employees that have bachelor's degrees on all exogenous firm characteristic variables along with the Hausman-type instrumental variable

[55] As shown in the Pearson correlation coefficient in table 1, the coefficients are small(<0.3).
[56] Mathematical details can be accessed through: https://www.nber.org/WNE/lect_6_controlfuncs.pdf
[57] In Blundell and Powell (2003), they release the fact the assumption of linear functional form in the first stage. They propose a complete non-parametric method. They also show that as long as the basic assumption regarding the relationship between the first stage residual and the error term of the linear parametric control function approach holds, this method should generate more consistent results when nonlinearity exists. Mathematical details can be found in their paper.
[58] Mathematical proofs and details can be found in Friedman(1991). MARS generally search for knots of each variable and estimate the coefficients of the interaction terms between variables before and after the knot. Then all terms can be added by in a linear form to generate the predictions for the dependent variable. Though all terms in MARS are combined in the linear form, it's still considered as a non-parametric method since the interpretation is difficult and meaningless.
[59] It's possible to determine those by using cross validation. However, Aydin (2007) empirically shows that the spline regression is better than kernel regression by using different valuation metrics.

either by using Ordinary Least Square (OLS) or using Multivariate Adaptive Regression Spline (MARS). For the control function approach with MARS residual, we estimated specification (4) for the first stage estimation.

$$bachratio_{it} = MARS(z_{it}) + v_{it} \tag{16}$$

Where MARS(.) is the Multivariate Adaptive Regression Spline function.

2, Compute the fitted value for the proportion of employees that have bachelor's degrees and compute the error between the fitted value and the actual value.

3, Regress dependent variables on all variables and the error term obtained from step 2, as specified in equation (15).

### 1.6 Baseline results

The estimated model of the baseline results presented in this section is specification (3). In the baseline result section, we also attach the results of estimations with a linear control function approach for the purpose of comparison. The main takeaways from the baseline results are 1, the marginal effects of increasing the proportion of employees with bachelor's degrees are always larger than 0 but not always statistically significant; 2, the polynomial terms are all significant so simple linear models are not appropriate; 3, the MARS control function approach fits the data the best compared to others.

The baseline results are presented in Table 3 (a) & (b)[60].

**Table 3(a) – Baseline results of regression analysis**

| | ***Baseline Results*** | | | | | |
|---|---|---|---|---|---|---|
| | *Dependent variable:* | | | | | |
| | *ROA* | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | ***OLS without control function*** | ***2-steps GMM with linear control function approach*** | ***OLS with linear control function approach*** | ***OLS with MARS control function approach*** | ***OLS with Kernel Regression control function approach*** | ***OLS with MARS control function approach*** |
| | **10.63***** | **5.21**** | **5.21**** | **18.22***** | **24.71***** | **11.75***** |
| *bachratio* | **(1.62)** | **(2.30)** | **(2.04)** | **(1.74)** | **(2.25)** | **(2.11)** |
| | -23.03*** | -22.60*** | -22.60*** | -21.30*** | -21.80*** | -9.05 |
| *bachratio-sq* | (5.54) | (5.54) | (3.70) | (5.50) | (5.48) | (6.25) |
| | 18.03*** | 17.73*** | 17.73*** | 16.56*** | 17.40*** | 14.25*** |
| *bachratio-cubic* | (5.17) | (5.16) | (3.13) | (5.09) | (5.11) | (5.32) |
| | -5.46*** | -5.39*** | -5.39*** | -5.66*** | -5.63*** | |
| *Log(asset)* | (0.44) | (0.08) | (0.51) | (0.08) | (0.09) | |
| | 0.46*** | 0.12 | 0.12 | 0.97*** | 1.42*** | |
| *Log(emp)* | (0.44) | (0.11) | (0.29) | (0.06) | (0.11) | |
| | **0.12***** | **0.20***** | **0.20***** | **-0.0247** | **-0.11***** | **-0.07*** |
| *Log(Avg_Salary)* | **(0.028)** | **(0.04)** | **(0.03)** | **(0.03)** | **(0.04)** | **(0.04)** |
| | -0.043*** | -0.043*** | -0.043*** | -0.040*** | -0.04*** | |
| *age* | (0.005) | (0.005) | (0.005) | (0.005) | (0.01) | |
| | 4.67*** | 4.80*** | 4.80*** | 4.59*** | 4.30*** | |
| *Log(exp)* | (0.08) | (0.09) | (0.03) | (0.08) | (0.09) | |
| | 45.98*** | 47.14*** | 47.14*** | 44.89*** | 0.33*** | |
| *markup* | (0.65) | (0.72) | (0.46) | (0.64) | (0.02) | |
| | 0.12*** | 0.05* | 0.05* | 0.23*** | 42.48*** | |
| *Log(Q)* | (0.15) | (0.03) | (0.03) | (0.16) | (0.73) | |
| | | 5.33*** | 5.33*** | -9.63*** | - 15.46*** | -9.52*** |
| $\hat{v}$ | | (1.54) | (1.59) | (0.72) | (1.59) | (0.59) |
| | 12.25*** | 13.01*** | 13.01*** | 10.63*** | 9.82*** | 3.94*** |
| *Constant* | (0.77) | (0.79) | (0.64) | (1.72) | (0.73) | (0.36) |
| *Adjusted R-sq/Obj-function value* | 0.68 | 0.0000118 | 0.68 | 0.69 | 0.68 | 0.05 |
| *N* | 13501 | 13501 | 13501 | 13501 | 13501 | 13501 |
| *Industry fixed effect* | Yes | Yes | Yes | Yes | Yes | No |
| *Year fixed effect* | Yes | Yes | Yes | Yes | Yes | No |
| *Province fixed effect* | Yes | Yes | Yes | Yes | Yes | No |

[60] For all tables in this thesis, the numbers in the parentheses are bootstrapping standard error unless otherwise specified.

**Table 3(b) – Fixed-effects estimation results**

| | ***Fixed-effects Model*** | | |
|---|---|---|---|
| | *Dependent variable:* | | |
| | *ROA* | | |
| | (1) | (2) | (3) |
| | ***FE with linear control function approach*** | ***FE with MARS control function approach*** | ***FE with MARS control function approach*** |
| *bachratio* | **16.59***** **(2.46)** | **12.05***** **(1.87)** | **7.74***** **(0.77)** |
| *bachratio-sq* | -15.28*** (5.07) | -12.93** (5.05) | |
| *bachratio-cubic* | 12.24*** (4.25) | 10.36** (4.24) | |
| *Log(asset)* | -4.84*** (0.08) | -4.95*** (0.09) | -4.94*** (0.08) |
| *Log(emp)* | 0.56*** (0.13) | 0.28*** (0.08) | 0.27 (0.76) |
| *Log(Avg_Salary)* | **-0.15**** **(0.04)** | **-0.08** **(0.03)** | **-0.08***** **(0.03)** |
| *age* | 0.03 (0.02) | -0.014 (0.016) | 0.19** (0.016) |
| *Log(exp)* | 4.39*** (0.08) | 4.62*** (0.07) | 4.62*** (0.72) |
| *markup* | 39.42*** (0.51) | 41.34*** (0.35) | 41.30*** (0.35) |
| *Log(Q)* | -0.14*** (0.05) | -0.19*** (0.07) | -0.20*** (0.05) |
| $\hat{v}$ | -10.68*** (1.69) | -7.14*** (0.71) | -7.21*** (0.71) |
| *Constant* | 11.77*** (1.62) | 13.00*** (1.62) | 13.21*** (0.16) |
| *R-sq* | 0.63 | 0.65 | 0.64 |
| *N* | 13501 | 13501 | 13501 |
| *Industry fixed effect* | No | No | No |
| *Year fixed effect* | Yes | Yes | Yes |
| *Province fixed effect* | No | No | No |

By looking at the coefficients of the term *bachratio*, we can easily see that no matter what kind of estimator we use to do the test, the polynomial terms of the proportion of employees with bachelor's degrees variable always significant impacts on ROA. The key insights provided by this table include the following: 1, By looking at the estimators generated from different estimation procedures, we can confirm the existence of the endogeneity; 2, By looking at the OLS estimation with MARS control function approach and the OLS estimation with linear control function approach, we can also confirm the existence of bias due to the non-linear structure and; 3, Comparing the estimation results generated from OLS and GMM, we can see the GMM is slightly more efficient, which implies the existence of heteroscedasticity. However, that doesn't change the results of the tests on the coefficients. A visualization of the marginal effect of the proportion of employees that have bachelor's degrees is attached in Figure 6.

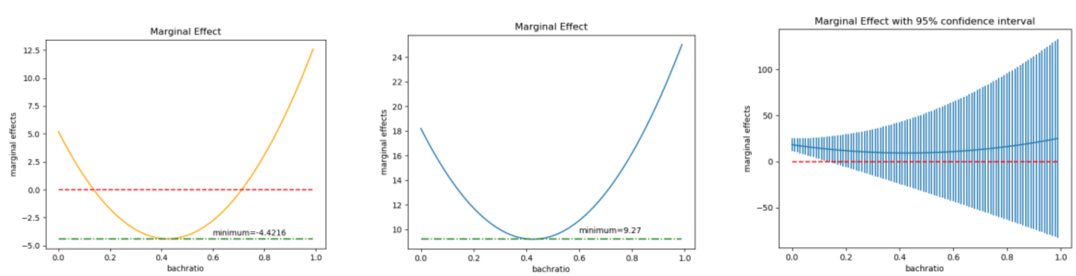


**Figure 6 – Visualization of marginal effects** [61]

As indicated by the estimation results and the visualization, the marginal effect from the proportion of employees that have bachelor's degrees is in quadratic form. If we look at the estimators generated from the OLS estimation with MARS control function,

[61] The method and proof for calculating the confidence interval for the marginal effect is attached in Appendix C.

the marginal effect equals to $\frac{\partial E[ROA]}{\partial bachratio} = [18.22 -$ $42.3\text{*bachratio}+49.68\text{*bachratio}^2]|_{bachratio=bachratio_{it}}$. At the average level of *bachratio*, which equals to 0.24, the marginal effect equals to 10.71, which means when the proportion of employees with bachelor's degrees increase by 1% for an average firm, the ROA is expected to increase by 0.107% in magnitude. By looking at the subgroups and calculating by using the same method, an average firm in the pharmaceutical industry is expected to gain an increase of 0.119% while an average firm in the IT industry is expected to gain 0.097% in the magnitude of ROA by increasing 1% in the proportion of employees with bachelor's degrees. However, this is not always statistically significant at the 95% confidence level. If we look at the right panel in Figure 3, we can see that only when bachratio is less than or equal to 0.1321,[62] the marginal effect is statistically significantly different from 0. That is being said, when the firm's bachratio is less than or equal to 0.1321 and the firm increases 1% in the proportion of employees holding at least bachelor's degree, the firm can experience an increase in ROA by $[18.22 -$ $42.3\text{*bachratio}+49.68\text{*bachratio}^2] * 0.01|_{bachratio=bachratio_{it}}$ percent in absolute magnitude.

By looking at the middle left panel in Figure 6 combining with the estimators obtained from the OLS estimation with linear control function, the marginal effect is non-linear, and equals to

[62] Another experiment shows that at the confidence level of 99%, only the firms that have a proportion of employees with bachelors' degrees lower than 0.0712 will gain significantly positive marginal effects.

$\frac{\partial E[ROA]}{\partial bachratio} = [5.21 \text{ -45.3*bachratio+53.19*bachratio}^2]|_{bachratio=bachratio_{it}}$. However, this effect can become 0 or even become negative when $bachratio$ increases.

That is to say, the marginal effect is only significantly positive within a certain range. What we can conclude so far is that the impact of aggregate education level on firm profit can be significantly positive but not always significantly positive. The firms having lower the proportion of employees that have bachelor's degrees (lower than 13.21 %) will most likely gain a positive increase in their performances if they hire more laborers with bachelor's degrees.

The other interesting thing shown from estimation results is the coefficients of the term Log (*Avg_Salary*) are always significant and positive when no control function is implemented and when the linear control function is implemented. When MARS control function approach is implemented, this term becomes negative or insignificant. This finding also confirms the existence of a nonlinear relationship between the exogenous variables and the endogenous variables. By using the linear form, it will still introduce bias to the estimation. In addition, I used the kernel regression first stage estimation to compute the first-stage error as a part of the control function approach. As indicated in Table 3 estimation (5), the estimators are larger in magnitudes than the estimators using the MARS first stage error when using the kernel regression for the first stage error computation. However, the R-sq is lower at the same time. That implies the kernel regression first-stage error might introduce more biases to the model and makes the fitting worse than the MARS first-stage error.

One of the concerns of mine for the empirical model in this paper is that I don't take the unobserved firm fixed effects into account, and I don't control the firm fixed effect. I don't control for the firm fixed effect is because there will be too many dummy variables to estimate. However, the unobserved firm fixed effect might also cause endogenous problems and biases. Thus, I also conduct a Fixed-effects estimation combining with the control function approach. The estimation results are shown in Table 3(b)[63]. Visualization is also attached in Figure 3 in Appendix B.

The linear fixed-effects estimators show that the marginal effect is supposed to be positive, which aligns with the estimation results by using the polynomial form. The other terms remain almost the same when getting rid of the polynomial terms, which indicates that the polynomial terms don't incur the multicollinearity issue.

As suggested by the results and the visualization, the positive within effects are still robust for a portion of firms with lower proportions of employees who are college degree holders. The magnitudes of coefficients are not much different from the coefficients in the baseline results. This confirms the robustness of the baseline results even though when conducting baseline estimations, no firm fixed effects are controlled.

These findings can't confirm our hypothesis H1 that the marginal effect from the proportion of employees that have bachelor's degrees on firm performance and profitability is always positive. However, we can conclude that the firms with a lower proportion of employees that have college degrees (<13%, as indicated in Figure 6) still can benefit from

[63] Please note that in Appendix B Table 4 the province fixed effect and industry fixed effect are not controlled since they are time-invariant and can't be identified in the fixed effect estimation. Though they can be recovered by the Hausman-Taylor method, it's not necessary and out of the cope of this paper.

hiring more employees with higher education degrees. However, this kind of benefit will decrease and become insignificant when they start hiring more and more new employees with bachelor's degrees. This finding, to a certain extent, justifies the Chinese government's policy of increasing enrolment quota for Chinses colleges: For the firms with few employees that graduated from college, they are anticipated to employ more laborers with college degrees. So, the supply needs to be guaranteed with the increasing enrolment.

Another interesting finding is the firm's wage level. The linear control function approach OLS results are telling that offering higher average salaries can increase the profitability and the performance of the firm. This aggregate-level finding basically aligns with the previous human capital research results. Murphy (1985), Leonard (1990) and other labor economics researchers find out that executive pay has a positive correlation with a firm's performance in terms of profits, return on assets, and market values. Possible reasons may be because the firm can attract more productive laborers by offering high salaries. Those laborers can create more values than the salaries that they are paid. Offering higher salaries can possibly guarantee the labor quality and attract more productive labors, and hence the firm's performance and profitability will be increased. However, this result is not robust: when conducting the same regression analysis with the MARS control function, this finding is overturned. Therefore, it can be inferred that the positive marginal effect is coming from the bias introduced by the non-linear problem.  Consequently, we can't draw any conclusion on the role of the wage level in firm performance. Further inquiries must be conducted before concluding.

To a certain extent, this paper confirms some previous findings: It shows that more employees with college degrees can be beneficial and help improve the performance of the firm has a low proportion of employees with college degrees. However, cautions will be needed when interpreting this finding since the marginal effect of adding more employees with college degrees is not always statistically significant and positive, especially for the firms that already have a high proportion of employees who have college degrees.

## 1.7 The Mechanism behind the Scene

### *1.7.1 Revenue vs. Expenses*

The baseline results above only give us a big picture of what kind of effect may the proportion of employees that have bachelor's degrees has on the firm's performance. Briefly, the previous section confirms the positive and significant impact on firms with a lower proportion of employees with college degrees. The last section shows that the cut-off value is 13.12% at the 95% confidence interval. However, it doesn't address the question of how. In this section, I will do an empirical test on hypothesis H3 to see whether the increase in the proportion of employees that have bachelor's degrees can generate comparative advantage in operating activities.

The specification for regression in this section is still specification (3). Since we can only conclude that firms having less than 13.12% of employees who graduate from college will benefit from hiring more college degree holders, I abstract a subsample for this section. The subsample is made up of all firm-year observations with bachratio<=0.1312. To test how employees with college degrees can influence the revenue and expenses of the firms, I run regressions on both dependent variables. The results are attached in Table 4 below:

**Table 4 – Revenue increase or expenditure decrease**

***Revenue v.s. Expenditure***

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | *Log(exp)* | | *Log(rev)* | |
| | (1) | (2) | (3) | (4) |
| | ***OLS with linear control function approach*** | ***OLS with MARS control function approach*** | ***OLS with linear control function approach*** | ***OLS with MARS control function approach*** |
| *bachratio* | **18.86***** **(3.02)** | **12.62***** **(3.73)** | **19.14**** **(3.04)** | **12.77***** **(3.74)** |
| *bachratio-sq* | -41.43 (44.30) | -113.78** (54.62) | -44.36 (44.46) | -115.96** (54.77) |
| *bachratio-cubic* | 94.88 (197.50) | 352.63 (243.52) | 102.45 (198.20) | 357.206 (245.14) |
| *Log(asset)* | 0.33*** (0.01) | 0.66*** (0.01) | 0.33*** (0.01) | 0.66*** (0.01) |
| *Log(Emp)* | 1.08*** (0.02) | 0.40*** (0.02) | 1.08*** (0.02) | 0.40*** (0.02) |
| *Log(Avg_Salary)* | **-0.14***** **(0.01)** | **0.09***** **(0.01)** | **-0.14***** **(0.01)** | **0.09***** **(0.01)** |
| *age* | -0.003*** (0.001) | -0.005*** (0.001) | -0.003*** (0.001) | -0.005*** (0.001) |
| *markup* | -5.08*** (0.07) | -3.33*** (0.08) | -3.69*** (0.08) | -1.94*** (0.08) |
| *Log(Q)* | 0.20*** (0.04) | 0.04*** (0.003) | 0.19*** (0.03) | 0.04*** (0.004) |
| $\hat{v}$ | -13.58*** (0.28) | -0.72*** (0.17) | -13.56*** (0.29) | -0.63*** (0.17) |
| *Constant* | -1.28*** (0.17) | 1.50*** (0.20) | -1.30*** (0.18) | 1.47*** (0.20) |
| *Adjusted R-sq* | 0.93 | 0.89 | 0.92 | 0.69 |
| *N* | 4318 | 4318 | 4318 | 4318 |
| *Industry fixed effect* | Yes | Yes | Yes | Yes |
| *Year fixed effect* | Yes | Yes | Yes | Yes |
| *Province fixed effect* | Yes | Yes | Yes | Yes |

Based on the estimation results in Table 4, the higher the overall education level is, the higher the total expenses it will incur for those firms. In the meanwhile, the total revenue goes up as well. Since the linear control function OLS has a much higher R-sq, it means the linear approach model fits the data better. By looking at the coefficients, the marginal effect from the employee education level is generally higher on revenue than on expenses. Therefore, it's legit to infer that the channel for the positive impact from the overall education level of employees is the increase in revenue, and the rise in the proportion of employees that have bachelor's degrees does bring competitive advantage to the firm when conducting operational activities.

Something is interesting regarding the econometric model that we use, the adjusted R-sq of the models are all above 0.8 and even close to 1. That means this model provides exceptionally strong explanatory power to the variations in revenue and expenditure of Chinese firms. This model should contribute the empirical corporate finance and empirical I/O research in the sense that so far, this model provides the most explanatory power to variations in Chinese firms' revenues and expenditures compared to previous literature.

### *1.7.2 Operating income vs. non-operating income*

Revenue is a vast concept and can be decomposed into non-operating revenue and operating revenue[64]. The former mainly refers to the income incurred through value changes of assets, arbitrages obtained from trading liquid assets, and other incomes coming from investment behaviours. The latter mainly refers to the revenue gained by providing products and services to customers within the range of the main business. Though we have learned that the positive effect from overall education level on firm performance is delivered through the increase of revenue, we still don't know whether the effect is coming

[64] Under both current IFRS and current GAAP standards, incomes can be decomposed into operational income and non-operational income.

from the increase in non-operational income or the increase in operating income. In order to address this confusion, I run regressions on non-operational income and operational income separately to test hypothesis H4.

In this part, since both types of incomes are subsets of revenue, I use the model from the previous section that is used to regress on the revenue. For the non-operational revenue part, I run regressions on both the log transform of non-operational revenue and value change income for the sake of robustness. For the operational revenue part, I run regressions on the log transform of operational revenue. Results are demonstrated in Table 5 below:

**Table 5 – Decomposition of the increase in revenue**

***Operational Income v.s. Non-operational Income***

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | OPI/Profit | | NOPI/Profit | ValueCh/Profit | |
| | (1) | (2) | (3) | (4) | (5) |
| | ***OLS with linear control function approach*** | ***OLS with MARS control function approach*** | ***OLS with linear control function approach*** | ***OLS with MARS control function approach*** | ***OLS with linear control function approach*** |
| | **5400.29**** | **4652.49*** | **-2686.05*** | **-2235.94** | **-2181.53***** |
| *bachratio* | **(2454.91)** | **(2470.39)** | **(1633.35)** | **(1641.78)** | **(1132.43)** |
| | -45489.90 | -54071.18 | 22629.5 | 27682.65 | 22230.08 |
| *bachratio-sq* | (35906.39) | (36097.08) | (23889.9) | (23989.56) | (16545.18) |
| | 170371.2 | 200887.1 | -89658.91 | -107611.1 | -97532.09*** |
| *bachratio-cubic* | (160055.6) | (160930) | (106491.1) | (106951.6) | (73744.05) |
| | -68.36*** | -29.01*** | 33.02*** | 9.50 * | 31.12*** |
| *Log(asset)* | (9.53) | (8.58) | (6.33) | (5.70) | (4.49) |
| | 130.63*** | 49;42*** | -70.42*** | -21.99 *** | -54.35*** |
| *Log(emp)* | (14.55) | (10.22) | (9.67) | (6.79) | (6.91) |
| | **-14.00***** | **13.64***** | **9.59**** | **-6.86**** | **5.99**** |
| *Log(Avg_Salary)* | **(6.33)** | **(5.05)** | **(4.20)** | **(3.35)** | **(2.95)** |
| | -3.30*** | -3.52*** | 0.97 * | 1.10 *** | 1.65*** |
| *age* | (0.69) | (0.89) | (0.59) | (0.59) | (0.42) |
| | 316.35*** | 524.59*** | -166.51*** | -290.22 *** | -94.09*** |
| *markup* | (53.84) | (53.25) | (40.44) | (35.39) | (28.27) |
| | 20.53*** | 2.05 | -4.76* | 6.24 *** | -13.59*** |
| *Log(Q)* | (4.14) | (2.75) | (2.51) | (1.83) | (2.15) |
| | -1612.40*** | -54.75*** | 955.94 *** | 41.03 | 694.10*** |
| $\hat{v}$ | (275.49) | (130.84) | (153.64) | (72.71) | (109.81) |
| | -384.31*** | 13.01*** | 164.97* | -29.96109*** | 266.31*** |
| *Constant* | (117.54) | (0.79) | (91.60) | (86.96) | (66.99) |
| *Adjusted R-sq* | 0.05 | 0.03 | 0.04 | 0.03 | 0.04 |
| *N* | 4317 | 4317 | 4317 | 4317 | 4317 |
| *Industry fixed effect* | Yes | Yes | Yes | Yes | Yes |
| *Year fixed effect* | Yes | Yes | Yes | Yes | Yes |
| *Province fixed effect* | Yes | Yes | Yes | Yes | Yes |

The implications of Table 5 are straightforward. The coefficients are supporting the story of increasing operational income: When regressing the non-operational revenue-to-profit ratio on *bachratio*, the coefficient is negative also not statistically significant. When regressing the operational revenue-to-profit ratio on *bachratio*, the coefficient is positive and significant[65]. That means the effect from the overall education level on operating income is positive, which implies that when more employees are holding college degrees, the non-operational income will get lower while the operational income will get higher. The estimators obtained from the regression on value change income/loss confirms this.

The findings stay aligned with the argument made by previous literature saying human capital may also help firms improve quality of service and hence, increase the firm performance. The empirical results in this section show that firms with more employees who have received higher education will be more focused on their primary business line and gain more income from their main business line and then experience an increase in operating income. Though this is an interesting finding, with the limited data I have, it's hard to test further what the reason is. Following the previous literature (e.g. Ittner and Larcker, 1998, and Ton, 2009), a possible explanation is that firms with more employees who receive college education have better service quality in their primary business line, and get reputation and customer satisfaction, which creates stable business relationships with their customers.

### *1.7.3 Different effects for different firms*

Based on the previous literature examining the labor market like Autor et al. (2007) and Autor et al. (2013), the needs for different types of skills vary across firms. Therefore, the importance of employee education should also vary across firms.

[65] At the confidence level of 90%.

To test H5 is true or not, we use the random coefficient model to do an estimation. The rationale behind the random coefficient model is that it assumes the marginal effects to be made up of two parts: the common effect shared by all individual firms and the variations among individuals($\beta_{1i} = \bar{\beta}_1 + V_{it}$). In order to simply the model and make it estimable, the following assumptions are made:

1, The variation $\boldsymbol{V_i}$ is uncorrelated with the endogenous variable $educ_{it}$.

2, $E[e_{it}|educ_{it}] = E[e_{it}|v_{it}] = \omega v_{it}$

3, $E[V_{it}|educ_{it}] = E[V_{it}|v_{it}] = \rho v_{it}$

4, I also assume even though the magnitude of the marginal effects may vary across firms, the curvature of the marginal effect remains the same for all firms. Therefore, the coefficients of the squared term and the cubic term will not have random components.

5, $E[V_{it}|e_{it}] = E[V_{it}]$ so that there's no need to construct the variance-covariance between them in a specific form.

Then specification (15) would become

$$y_{it} = a + \beta_1 educ_{it} + \beta_2 {educ_{it}}^2 + \beta_3 {educ_{it}}^3 + \omega\hat{v}_{it} + \rho\hat{v}_{it} * educ_{it} + \boldsymbol{\delta X_{it}} + \boldsymbol{\theta D_{it}} + e_{it} \qquad (16)$$

The results of the estimation on specification (16) are shown in Table 6.

**Table 6 – Random Coefficient Model[66]**

[66] The variable *bachratio* is the aggregated education level of employees measured by the proportion of proportion of employees holding at bachelor's degree or above. Log(exp) is the natural log transformation of expenditure. Log(rev) is

**_Random Coefficient Model_**

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | *ROA* | | *ROA* | |
| | (1) | | (2) | |
| | ***OLS with linear control function approach*** | ∑ | ***OLS with MARS control function approach*** | ∑ |
| *bachratio* | **7.55**<br>**(1.45)** | **44.32*****<br>**(2.80)** | **8.63*****<br>**(0.54)** | **43.01*****<br>**(2.68)** |
| *bachratio-sq* | 103.13***<br>(4.67) | | 91.98***<br>(4.64) | |
| *bachratio-cubic* | -127.99***<br>(7.99) | | -110.90***<br>(7.85) | |
| *Log(asset)* | -5.23***<br>(0.06) | | -5.44***<br>(0.06) | |
| *Log(emp)* | 0.53***<br>(0.11) | | 0.88***<br>(0.06) | |
| *Log(Avg_Salary)* | **0.07****<br>**(0.035)** | | **-0.003**<br>**(0.027)** | |
| *age* | -0.042***<br>(0.01) | | -0.048***<br>(0.01) | |
| *Log(exp)* | 4.50***<br>(0.07) | | 4.53***<br>(0.06) | |
| *markup* | 43.53***<br>(0.45) | | 43.49***<br>(0.31) | |
| *Log(Q)* | 0.003***<br>(0.03) | | 0.025***<br>(0.02) | |
| $\hat{v}$ | -6.06***<br>(1.44) | | -8.26***<br>(0.62) | |
| $\hat{v} * bachratio$ | 0.55<br>(1.12) | | 0.30<br>(1.16) | |
| *Constant* | 21.07***<br>(0.95) | | 20.13***<br>(0.92) | |
| *Log-likelihood* | -31787.49 | | -31679.42 | |
| *N* | 13199 | | 13199 | |

---

the $\hat{v}$ is the residual generated by using the control function approach. ∑ denotes the variance-covariance matrix of the random coefficients.

Some firms can enjoy a marginal effect of up to 0.37% increase in ROA when they increase the proportion of employees with college degrees by 1%.[67] Therefore, the results confirm the argument proposed in Autor et al. (2007) and Autor et al. (2013): The needs for different types of skills vary across firms, and the importance of employee education also varies across firms. The marginal effects brought by employees with higher education differ from one firm to firm.

## 1.8 Robustness checks

### *1.8.1 Measurement of firm performance*

The measurement of firm performance has always been an issue in corporate finance research. As mentioned earlier, the mainstream scholars in the field of corporate finance now agree that ROA, ROE, and Tobin's Q are commonly used measurements of firm performance. This section is following the idea of the robustness check in Giroud and Mueller (2010). As emphasized, ROA usually accounts for profitability, and profitability is the most important feature of a firm since the ultimate goal of a firm is to maximize the profit so we choose ROA as the dependent variable for the baseline regression. Therefore, the robustness checks will be performed on ROE and ROIC. The ROE ratio is another measurement that is widely accepted as the performance measurements of firms. The ROA usually doesn't take the financial leverage into account while the ROE ratio usually can measure how well the firm is generating income on the equity when financial leverage is taken into account. Another commonly used measurement is Tobin's Q, which measures

[67] As indicated by Figure 4, the maximum of the marginal effect is the firms with 2 standard deviations above the mean level. When the proportion of employees with college degree approach to 1, they have a maximum marginal effect, which is about 37 in magnitude.

whether the investors appreciate the firm and whether its value is appraising. The robustness check results are attached in Appendix A Table 2. Since in the baseline results section, it's shown that the MARS control function approach has a better fit for the data, and corrects the bias on both the coefficient of employee education and the coefficient of wage level, robustness checks are conducted only by using the OLS with MARS control function approach in this section. As shown from the estimation results and the visualization, our conclusion remains true: The marginal effects on firm performance from the employees holding college degrees are non-linear, and also not for all firms the effect is significantly positive. Only those firms with a relatively lower proportion of employees holding college degrees can experience significant positive effects.

### *1.8.2 Stock performance*

As mentioned previously, the measurements of firm performance commonly used in the field of corporate finance and empirical I/O are all accounting measurements. Some may be concerned that these accounting measurements can be manipulated by playing with the accounting criteria and managing the balance sheets or income statements. Thus, some scholars (such as Giroud and Mueller, 2010) also look at the cumulative abnormal returns of the stocks of the firms to see whether the findings are robust or not. Here I also use this method to test hypothesis H6 to ensure the robustness of our conclusions. I use the classical CAPM model[68] to estimate the predicted return. Then I compute the cumulative abnormal returns by deducting the predicted return from the realized return. Then I test whether the

[68] The CAPM model is a classic model that tries to find the correlation between return and risks. The original CAPM model only contains the pricing load of risk. However, recently the three factors model, four factors model and five factors model also propose that the momentum, firm size, market-to-book ratio should also be factors that provide essential explanatory power to the variation in asset price. Pástor et al. (2000) proves that in reality, these three models make no difference when investors are making investments by looking at the abnormal returns.

CAR of the portfolio which contains the stocks of the firms with different overall education levels are the same or not. The first set of test results shows that the portfolio made up of firms with high overall education levels (bachratio>=0.24)[69] has a higher mean of CAR than the portfolio made up of firms with low overall education levels (bachratio<0.24). The student-t test results show that the difference is significant. To sum up, this is confirming the findings of previous sections. The test results are shown in Table 5.

However, since the firms might have different characteristics that may drive their stock prices to be different, dividing the firms into the two groups by using the mean is not adequate because there might be other factors that may influence the CARs[70]. Therefore, we use the clustering technique from unsupervised learning to cluster them into groups. The essential spirit of clustering is to make the comparable data points all into the same group and match with each other. We chose the Multinomial Normal Mixture (MNM) model for clustering since we don't have to decide the number of clusters. Instead, it will determine the cluster number by maximizing the log-likelihood71. We use the subsample of the year 2016 to perform this clustering since the clustering requires the same time dimension. The clustering results are shown in Figure 2 in Appendix A.

As indicated in the results in Table 3 Panel A in Appendix A, interestingly, the firms with an above-average proportion of employees with bachelor's degrees have significantly higher CARs than the firms with a below-average proportion of employees with bachelor's degrees. By looking at the test results on the clustering groups, those results

[69] The mean of the bachratio of all firms is 0.24.
[70] Because the time dimension of the data is year and the stock price reported Is the price of the adjusted close price of last trading day in the year, the time dimension of the CAR here is the CAR of a specific year assuming the abnormal return of the last trading day of the previous year is 0.
[71] Details and mathematic proofs can be found in Vermunt and Magidson (2000).

essentially confirms the findings in the baseline result in the sense that both of them shows firms with proportion of employees with college degrees larger than 0.131272 are likely to have lower CARs than the shows firms with percentage of employees with bachelor's degrees smaller than 0.1312. That also confirms our findings: 1, For firms having proportions of employees with bachelor's degrees larger than 0.1312, the marginal benefits for them to hire more college degree holders will be statistically indifferent from 0 while firms proportions of employees with bachelor's degrees smaller than 0.1312 can experience significantly positive benefits with such hiring; 2, The marginal effects from hiring college degree holders vary across firms even though the firms have similar characteristics.

### 1.9 Management and policy implications

As discussed, the key findings of this paper provide management and policy implications, and those implications include the following:

1, Not all Chinese public firms will gain performance improvement by hiring more college graduates. If the firm currently has a low proportion of employees with bachelor's degrees, more precisely lower than 13%, then hiring more college grads might benefit the firm. More specifically, according to the data, firms in the retail industry and hotel & traveling services industry should expect to gain significant improvements by hiring more college graduates.

[72] The critical value 0.1312 is determined by the baseline estimation results: For firms have a proportion of employees with college degrees larger than 0.1312, the marginal benefits for them to hire more college degree holders will be statistically indifferent from 0.

2, The firms with a low proportion of employees with bachelor's degrees should gain more operating income through hiring more college grads. Hiring more employees with bachelor’s degrees can improve the income gained from the firm’s primary product/service line.

3, The policy of enlargement of college enrolment in China can be beneficial for Chinese firms in the sense that currently, many public firms still have low proportions of employees with bachelor’s degrees. Those firms, more precisely, the firms in the retail industry and hotel & traveling services industry, can benefit from hiring more college graduates, which creates a higher demand for college graduates in the labour market.

### 1.10 Concluding remarks

In this paper, we investigate a long-argued issue about whether hiring people with college degrees will improve the performance of a firm. On the one hand, some economist argues that hiring people with higher education levels leads to higher wage costs. On the other hand, some managers and economists say that people with higher education can increase the overall productivity of the firm. The data analysis based on the data of Chinese public firms supports the hypothesis that hiring people with higher education levels, especially with bachelor’s degrees will improve the performance. However, that’s only true for the firms with a relatively low proportion of employees who are college degree holders. For the firms with a relatively high proportion of employees who have college degrees, the effect is not significant and can be negative at some points. Therefore, no general conclusion can be made for all firms.

This paper also tries to explore where these impacts are from and how they are delivered. Therefore, I also test whether these effects are coming from the reduction of costs or the increments in income. It turns out that the improvement in firm performance is coming from the increase in operational income, which implies that the employees who have received higher education are more likely to make the firm more focused on its main business line and obtain more income from the main business line. In addition, the mixed-effect model confirms that the effect varies across firms and it's not unified.

The robustness checks confirm the baseline findings in this paper. The findings in this paper contribute to the literature of human resources management and labor economics in the sense that this paper provides a brand-new perspective on how the education of employees can influence the firm's performance. This paper may also contribute to the econometrics literature because the non-linear approach combined with machine learning techniques contributes to the empirical literature applying machine learning to managerial and policy-relevant issues.

In the meanwhile, there are some issues that this paper that can't be or wasn't addressed: Due to the limitation of the data, we can't control the propensity to IPO, which may create the selection biases for the estimators based on the data of public firms and the conclusion probably can't be generalized to the non-public firms. Because of the lack of individual-level data, there is no way for me to do structural estimation by considering the simultaneous decisions of firms and individual laborers. Besides, the industry distribution of public firms in the sample is significantly different from the industry distribution of all firms in China. Therefore, although providing important insights on the relationship between firm performance and a college-educated workforce, without a more representative sample of firms, one must be cautious in generalizing the conclusions in this paper to the private firms.

# CHAPTER 2. AN INVESTIGATION ON THE DEMAND IN THE LOAN MARKET: A STRUCTURAL MODEL OF CUSTOMER CHOICE

## 2.1 Introduction

This paper explores the key determinants that definite the market shares of banks within the loan market, and what kind of impacts there will be on the market structure if the Federal Reserve Bank changes the base interest rate. More precisely, this paper tries to provide answers to the following questions; 1, What are the key factors that determine the demand for a loan from a commercial bank? 2, Is interest rate a key factor in the competition? 3, If the answer to the previous question is yes, then does the monetary policy that changes interest rates reshape the loan market structure, and how will the monetary policy have impacts on the market structure?

The findings of my analysis include the following:

1) I find out that the ease of access to service, the service quality, and the interest rates are the key factors that determine the choice of customers. Based on the BLP[73] estimations and the 2SLS estimations for the demand curve, the ease of access to service bank, which is measured by intangible assets per employee and staff expenses per employee and number of employees has significant positive impacts on the share/demand of the loans offered by a bank. The service quality of a bank, which is measured by staff

[73] The BLP estimation refers to the estimation algorithm proposed by Berry et al. (1995).

expenses per employee, has significant positive impacts on the share/demand of the loans offered by a bank as well. The interest rate has an economically and statistically negative impact on the share/demand for loans provided by a bank.

2) the counterfactual experiment on interest rate changes shows that the small banks are more sensitive to the interest rate adjustments, and increases in interest rate can squeeze out the small banks from the loan market. The experiment shows that largest banks always have lower own--interest rate elasticities than smallest banks and that when the Fed increases one base point of the base interest rate, 4 out of 5 largest banks in the sample will lose less than 0.015% of the share while 4 out of 5 smaller banks in the sample will lose more than 0.015% of the share. This is the first analysis that empirically explains why that smaller banks are losing more than larger banks and hence, smaller banks tend to use more derivatives to hedge the interest rate risks as indicated by Fraser (2002).

3) banks with higher market power intend to be risk-averse. The analysis shows that the banks with a higher market power, measured by both HHI and own-interest rate elasticity, tend to have lower loan loss ratio and subordinated debt asset to total asset ratio. The positive correlation between the elasticity and the loan loss ratio and the positive correlation between the elasticity and the subordinated debt asset to total asset ratio are significant. This conclusion is perfectly aligning with the conclusion in Salas and Saurina (2003) though the dataset that they use is the Spanish bank dataset.

These results are crucial for two reasons: 1, The U.S. banking industry has been a key component in the economic system. Based on the data from The Bureau of Economic Analysis (BEA), the direct contribution from financial institutions on the whole economy is about 7 percent in 2018[74]. Understanding the factors that can shape the market structure of the lending market can provide policy implications to the monetary policymakers and enhance the efficacy of the polices made: Monetary policymakers can anticipate and quantify the impacts from an adjustment on interest rates with the loan market demand parameters, and can also predict the market structure changes with those parameters and evaluate the adjustments on interest rates accordingly, which may help policymakers avoid the undesired structural changes within the financial market effectively. 2, Consumer behaviors have also been discussed a lot[75], and it's always the center of industry study because the preference and demand of the customers are the foundations for industry analysis in terms of profit margin, return on equity, and market value estimation[76]. Therefore, knowing customers' preferences over banks will enhance the understanding of the competition within the banking industry, and will provide marketing or even strategic level insights for the bank managers.

During the last forty years, econometricians and macroeconomists developed methodologies to estimate consumer's demands conditioned on preferences. By leveraging these tools, the analysis for consumer choices over banks becomes possible[77]. This paper contributes to that literature by showing the micro-econometric techniques, namely the BLP demand estimation, can be applied to financial markets. Another contribution of this

---

[74] Based on the *Gross Domestic Product by Industry: Fourth Quarter and Annual 2018* published by BEA on April 19th, 2019. The full report can be assessed through https://www.bea.gov/system/files/2019-04/gdpind418_0.pdf. This number is measured by using output (GDP) contribution.
[75] In past 30 years, numerous researchers conduct consumer behavior research using either structural estimation or reduce-form methods to explore the consumer preferences in different good/service markets. Section II will review some closely related works.
[76] For instance, Nevo (2001) estimates the consumer preference parameters and conducted counterfactual experiments to evaluate the market share and possible revenue changes of the cereal manufacturers. Berry, Levinsohn and Pakes (1995) also estimates the consumer preference parameters over cars and conducted experiment to estimate the changes of profit margins of the automobile manufacturers.
[77] Appendix B gives a detailed review of the techniques.

paper is the confirmation that the price factor, the loan interest rate, and the service quality are factors that determine the market share. This research also uses counterfactual experiments to show that monetary policies that adjust the interest rates may have unexpected impacts on the structure of the commercial banking industry. Increasing interest rates may squeeze out the small banks from the loan market.

In addition to the above contributions, this paper also contributes to two sets of literature: 1, this paper confirms the finding of the survey conducted by Boyd et al. (1994) that reputation and interest rate are the most important elements to the U.S. customers when they are making choices over banks. 2, this paper contributes to the financial institution literature in the sense that it examines the loan market from a micro-econometric perspective though previous researchers, like Kim et al. (2003), Shy (2002) and Ho (2015) only explore the depositary sectors[78].

### *2.1.1 Institutional background*

#### 2.1.1.1 Banks in the U.S.

The U.S. banking industry has been a critical component in the economic system of the U.S. for centuries though some researchers argue that the importance of commercial banks is diminishing over time[79] (Kaufmann, 1993). Nowadays, banks are not only providing basic depository or lending services but also offer other kinds of financing

---

[78] Details can be found in the literature review section regarding the research of the depositary markets.

[79] Berger et al (1995) summarizes the transformation and transition history of the banking industry in the united states. By looking at the development and economic contribution of banks, they assert that "the banking industry is an integral part of U.S. Economy". Though Kaufmann (1993) argues that other intermediate financial institutions are stealing market shares in the loan market from commercial banks, the data shows that commercial banks still hold more than 70% of the loan outstanding nationwide.

services, like portfolio management and financial advisory, that outreach almost every facet of the economic life of an individual. The banking industry also acts as a practitioner of macroeconomic policies in the sense that it executes monetary policies made by the central banks. Therefore, the banking industry has frequently been the target of policy regulations[80]. Though banks are more diverse in the financing services that they provide, the most critical functions of commercial banks are still to provide loans and help individual or enterprises to finance their needs[81].

#### 2.1.1.2 Overview of the banking industry and the loan market

Based on the FDIC report, the total industry assets of the banking industry hit $18.09 trillion by March 2019[82]. According to Reuter's business news published on Tuesday, Feb. 17th, 2017, the net income after tax and interest expenses of the banking industry is $171 billion in 2016, an all-time high.[83] Compared to the data of the year 2015, there is a 4.9% increase in net income for the whole industry, which is almost equivalent to the GDP value of Vermont State in 2016.[84] There are currently 19,821 active financial institutions with depositary and loaning businesses having a business operating on the U.S.

---

[80] The development and functional transformation of banks are elaborated in Berger et al (1995). After the deregulation of banks over the late 1980s and early 1990s, banks became more diverse of business lines. Those new business lines include investment banking businesses, off-balance-sheet businesses (i.e. derivatives), insurance and risk management businesses, asset management and financial advisory businesses and other businesses that outreach almost all edges of economy. More details can be found in Berger et al (1995). Berger et al (1995) also list some others numerous capital regulations.

[81] See Kaufmann (1993). Commercial banks hold more than 70% of the outstanding loan in the U.S. Also, commercial banks have the highest proportion of loans in total assets (on average about 10%) compared to other intermediate financial institutions.

[82] FDIC Quarterly report of industry trend which can be accessed through: https://www.fdic.gov/bank/statistical/stats/2019mar/industry.pdf

[83] The numbers in the article are obtained from the FDIC report which Can be accessed through: http://www.reuters.com/article/us-usa-banks-fdic-idUSKBN1671V7. Their news report is based on the Federal Deposit Insurance Corp figures, which are not inflation-adjusted.

[84] Based on the data collected by The Bureau of Economic Analysis of the U.S. The by- industry GDP data can be accessed though: https://www.bea.gov/data/gdp/gdp-industry.

domain.[85] There are roughly thirteen types of specialization among those financial institutions, including Specialized governmental credit institutions, Bank holdings & Holding companies, Commercial banks, Real Estate & Mortgage banks, Investment banks, Finance companies, Securities firms, Investment & Trust corporations, Clearing & Custody institutions, Cooperative banks, Private banking/Asset management companies, Savings banks, and Central banks. Among those institutions, commercial banks and bank holding companies make up the 75% percent and other sorts of financial institutions altogether make up the 25% percent[86]. Figure 7 (a) shows the distribution of bank specifications.

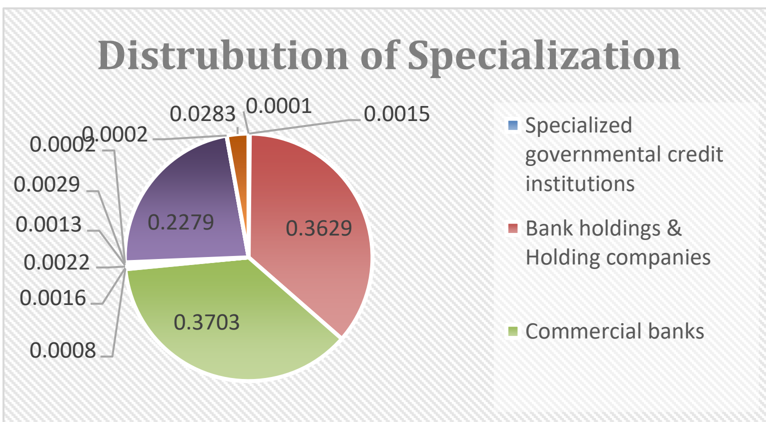


**Figure 7 (a) – Specialization distribution of the financial institution[87]**

As indicated by Figure 7 (b), the banking industry has been through a downfall during the financial crisis over the 2008-2011 period. During those years, the total net income dropped drastically. As shown in Figure 7 (b), the return on asset (ROA) and return

[85] Based on the data obtained from Bankscope Database. More details about this databased and dataset will be discussed in the data section.
[86] This information is generated based on the specification category in Bank Scope Database. More details about this databased and dataset will be discussed in the data section.
[87] Based on the data obtained from the Bankscope Database.

on equity (ROE) ratios have dropped tremendously. The ROE ratio drops from well above 10% to nearly 0% during 2008. During 2009 and 2010, both ratios increased drastically because of the macroeconomic policies and the interference of FDIC. Since the year 2012, the industrial ROA and ROE ratios both have been increasing steadily, and the ROE ratio remains on the level of close to 10%.

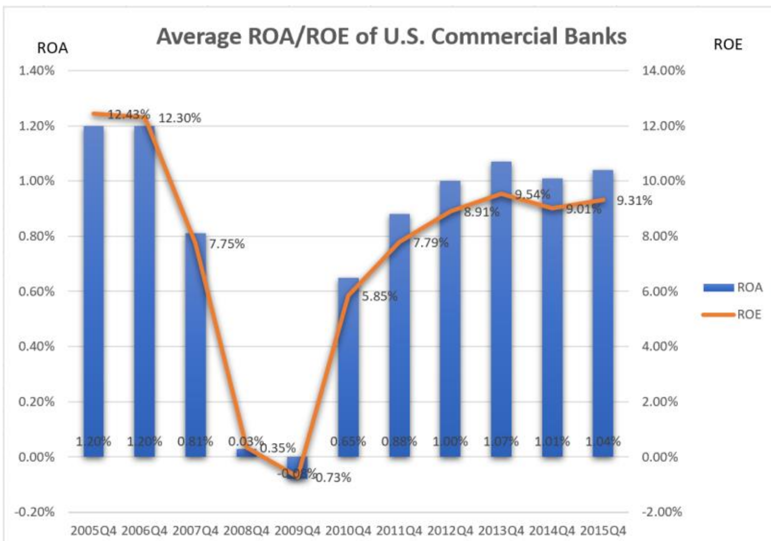


**Figure 7 (b) – ROA and ROE of commercial banks from 2005 to 2015[88]**

Based on the data published by The Bureau of Economic Analysis (BEA), the direct contribution from financial institutions on the whole economy is about 7 percent in 201889. Unfortunately, there’s no objective way to measure the indirect contribution, but

[88] Data obtained from FDIC quarterly reports from year 2005 Q4 to 2015 Q4. All reports can be accessed through https://www.fdic.gov/bank/analytical/quarterly/
[89] Based on the *Gross Domestic Product by Industry: Fourth Quarter and Annual 2018* published by BEA on April 19th, 2019. The full report can be assessed through https://www.bea.gov/system/files/2019-04/gdpind418_0.pdf. This number is measured by using output (GDP) contribution.

some economists argue that the total contribution from this sector should count for 8% or more of the economic growth.[90]

The loan market is enormous and diverse: Based on the FDIC annual report, the total outstanding loan value of individual loans for education (student loans), auto loans have reached a book value of 152 billion dollars in the U.S. On average, every U.S. resident has a balance of around 780 dollars of loan to pay and this number is net of real estate mortgage, commercial/business loans[91]. Reflecting its overall scale, the loan market attracts more than 1000 foreign commercial banks[92] to offer loan services in the U.S. territory.

Berger et al. (2004) and Cetorelli and Strahan (2006) find out that after the deregulation in the early 1990s, the bank industry became more diverse in terms of business as well as more locally concentrated. Though Shaffer (1989) previously found that the banking industry was trending perfectly competitive before the deregulation. As mentioned by Berger et al. (2004), Cetorelli and Strahan (2006) and Berger et al. (1995), after deregulation in the banking industry, banks were more concentrated and tried to make their services more differentiated to get higher concentration and market power, [93] and the whole market structure is moving towards a monopolist competition. Thereby, the BLP framework and the discrete choice framework should be the appropriate models that can provide answers to the question of what factors determine the demand for loans. The reasons for the adoption of the BLP framework in this paper include the following: 1, The

[90] There is a discussion about his issue: What is the contribution of the banking sector to the GDP? It can be accessed through https://www.quora.com/What-is-the-contribution-of-the-banking-sector-to-the-GDP.
[91] FDIC Quarterly report of industry trend which can be accessed through: https://www.fdic.gov/bank/statistical/stats/2019mar/industry.pdf
[92] Based on the data provided by Bankscope database. Details of this database will be elaborated in the section 3.
[93] For instance, Capital One is try to make their services more advanced by introducing more high level technologies which can facilitate their customer service. Details can be obtained from the official website of Capital One regarding their technological advances: https://www.capitalone.com/tech.

lack of individual choice data; 2, The non-linear price endogeneity problem in the demand estimation; 3, The BLP framework relaxes the unrealistic assumption of symmetric cross-price elasticities in standard discrete choice models. More details regarding the BLP framework will be discussed in section 2.2 and section 2.4.

## 2.2 Literature review

### *2.2.1 Brief view of related literature*

This paper mainly draws literature from three streams of works: 1, works in the field of industrial organization using structural models (e.g. the BLP framework) to estimate the demand for other industries; 2, determinants of customers' choices/preferences over bank; and 3, works using structural models to estimate customers' choices/preferences over the bank services.

### *2.2.2 Structural models and demand estimation for other industries*

The first stream of literature that I refer to include the empirical industrial organization works that use structural models to estimate demand in other sectors. Hausman et al. (1994) estimate the bear market demand by using the structural model. He follows the basic discrete choice framework to do a structural estimation. He estimates the elasticity of price and the impacts from other factors, such as weather, on demands. David (2006) uses the structural estimation method to estimate the demand in the theater industry. He not only estimates the price elasticity and the impacts from schedules of shows, but he also conducts a scenario analysis on the revenue of each theater based on the estimated preference parameters. Bresnahan et al. (1996) use the structural model to estimate the demand in the PC market for different brands. They particularly leveraged the nested structural model and relax the Independence and Irrelevant Alternative (IIA) assumption in the primitive multinomial logit model. Sudhir (2001) uses the standard BLP-type model to estimate the price elasticities and profit margins of the manufacturers in the yogurt and peanut butter market. Xu et al. (2016) use the structural estimation method, precisely the BLP, and Nevo methods, to estimate the impact of the online review on the demand of healthcare provided by each doctor. They conduct a sentimental analysis and text mining of the review. By using that information and the structural estimation method, they find out that the online review has significant impacts on the bookings of the doctors. There's also a substitution effect between the price and the review based on their counterfactual experiments. This set of literature provides insights on how to apply the BLP and Nevo's frameworks to a specific product/service market. However, only a few researchers look into the financial service market with these methodologies.

### *2.2.3 Determinants of customers' choices over banks.*

I also investigated the management and marketing literature regarding the determinants of customers' choices over banks. Boyd et al. (1994) firstly investigate the preferences of the consumers in the U.S. over financial services. They find out that reputation, helpfulness of staff, the efficiency of the service, interest rate and service charges, and depositary interests are the most important elements to the U.S. customers when choosing a financial institution regardless of age, marital status, income level, and household components. Zineldin (1996) has investigated the preference of customers of banks by using questionnaires. Based on his sample of Swedish customers of Swedish banks, he calculates the importance of the characteristics of banks and lists the most important factors that customers consider when choosing over banks. He finds out that the price and cost, helpfulness of the staff, service efficiency, and availability are the key elements that lead to a decision. Kennington (1996) also conducts similar research by using the Poland sample, he finds out that most of the respondents consider service quality, accessibility, and costs. His findings mostly align with Zineldin (1996). Phuong Ta et al. (2000) also conduct a hierarchical analysis by using the sample of Singapore. Their findings mostly concord to Kennington (1996) and Zineldin (1996). In addition, they find out that the overall reputation of the firm is also a key element in decision making. Kaynak et al. (1992) also got similar results by looking at the preference of Hong Kong residents over banks and find reliability, and risk is a factor that customers consider. Almossawi (2001) conducts an investigation regarding the preferences of college students over banks in Bahrain. He finds out that technology and reputation are the most important determents when the students are choosing a bank. He also finds out that convenience and service quality are also key elements. The customer choice factors utilized in the structural estimation model in this paper are selected based upon the results or findings from this literature.

### *2.2.4 Deposit markets and loan markets*

The number of previous research on loan markets is quite limited in both the industrial organization field and the field of finance. The most cited paper is Kim et al. (2003)'s paper. That paper focuses on the role played by switching costs in the consumer choices and market shares of banks in the Norwegian loan market. They derived a structural model to capture the oligopolistic structure of the Norwegian loan market and discuss how switching costs can affect the market share of the banks. Though this paper provides an estimation framework for the banking industry, due to the difference between the structure of the U.S banking industry and the Norwegian banking industry, the framework that they propose can't be applied to the U.S. banking industry: The American banks are fairly diverse in terms of service quality, size, IT research and adoption. The number of banks in the U.S is large. There are more than 18000 financial institutions and around 6000 commercial banks actively operating in the U.S. Due to the huge differences of characteristics between banks and the differentiation among banks, the loan services sector is no longer a homogenous-good market. Thereby, the U.S. banking industry and the loan service market is more like a monopolist competition structure, and this paper's framework doesn't apply to the U.S. loan market.

Several other papers look at the competition and market share problems in the banking industry. Shy (2002) and Ho (2015) both looked at the relationship between switching costs and the deposit market share of banks. Both of them leveraged the structural model that they derived by themselves. The key difference between their works is that Shy (2002) models the deposit as homogenous good while Ho (2015) models the deposit as a differentiated and durable good. The estimation procedure that is used by Ho (2015) is quite similar to a standard BLP, and hence fairly identical to the one in this paper. Besides, Ho (2015) provides some customer preference variables that may also be relevant to the choice of loans. However, Ho (2015) doesn't look at the borrowing side of the business— the loan market. In this paper, the framework that is specifically for the demand estimation in the market of loan.

## 2.3 Conceptual framework

### *2.3.1 Random utility and consumer choice*

I follow the standard random coefficient choice model framework. I observe T markets (T=1, 2, 3, 4….), each with hypothetically I[94] consumers and j (j=$J_T$) banks that vary by market. Furthermore, I assume each bank is captured by K characteristics. In typical choice models, a market is defined as a "city-time" combination (e.g. Berry et al., 1995, Nevo, 2001, and Ghose et al., 2012).

The utility of customer i from choosing bank j in market t, is denoted as $u_{itj}$, which is a function of observed and unobserved (by the econometrician) bank characteristics,

$x_{jt}$ and $\vartheta_{jt}$, respectively; and the market-specific characteristic, $z_t$. In my model, the market-specific characteristic is also important since it captures the systematic changes in market demand over time. As discussed in Section 1, the bank industry experienced fluctuations during my sample period. Computing the elasticities without controlling for the time fixed effect will lead to biased estimation because the unobserved year fixed effect may be correlated with bank characteristics. Traditional BLP-type models tend to focus more on product characteristics. Thus, the utility for consumer i from choosing bank j in market t can be represented as the following model:

$$u_{itj} = \beta_i x_{1tj} + \gamma x_{2tj} + \rho z_t + \vartheta_{jt} + \varepsilon_{ijt} \tag{1}$$

[94] In order to improve the efficiency of the estimation, Halton draw numbers usually is set at 1000.

where $x_{1jt}$ denotes the observed bank characteristics that customers value but the evaluations depend on the individual perspectives, for instance, the reputation of the bank $x_{2jt}$ denotes the characteristics that the customers universally consider regardless of individual views, for instance, the service quality.

Next, I also consider how consumer preferences vary as a function of individual characteristics. My approach here mainly follows Nevo (2001). The intuition behind the simulation is to capture the heterogeneous utility gain obtained by customers due to the heterogeneity in their preferences. Assuming normal distribution over of the customer preference, we have

$$\beta = \bar{\beta} + \sum v_i \, , v_i \sim P \; (\mu, \boldsymbol{\sigma}) \qquad (2)^{95}$$

where $P_v(\mu, \boldsymbol{\sigma})$ is a multivariate normal distribution of the demographic characteristics with mean vector of $\mu$ and variance-covariance matrix of $\boldsymbol{\sigma}$.

Thus, combing (1) and (2) we will get

$$u_{itj} = \delta\left(x_{1jt}, x_{2jt}, z_t, \vartheta_{jt}; \overline{\beta}, \gamma, \rho\right) + \mu \; (x_{1jt}, v_i; \Sigma) + \varepsilon_{ijt} \qquad (3)$$

where $\delta_{jt} = \bar{\beta}x_{1jt} + \gamma x_{2jt} + \rho z_t + \vartheta_{jt}$ is the mean utility and $\mu = \left[x_{1jt}\right]' * \sum v_i$, represents the deviation from the mean, $\sum is\ the\ vector\ of\ deviation\ from\ the\ mean.$ Customers are choosing the banks which

[95] In Berry et al. (1995), the distribution used is the pdf of income distribution. P(.) means the the distribution is a function of some parameters, for instance the income. This notation is following BLP (1995).

give them the highest utility, which defines the choice set of unobserved variables that lead to the choice of bank j:

$$A_{jt} = \{(v_i, \varepsilon_{ijt}) | u_{itj} > u_{itl}, \forall l = 0,1,2 \dots . J\} \tag{4}$$

Additionally, assuming ties occur with zero probability, the market share of the j-th bank as a function of the mean utility levels of all the $J_t$ + 1 choices (including outside good[96]), given the parameters, is

$$\mathrm{S}\ (x_{1jt}, x_{2jt}, z_t, \vartheta_{jt}, v_i; \overline{\beta}, \gamma, \rho, \textstyle\sum) = \int_{Ajt} d\, P(v_i) P\ (\varepsilon\ ) \tag{5}$$

As proposed by Berry et al (1995), given the computation of derivatives of shares, and to allow for interactions between individual consumer and specific characteristics, specification (2) is nested into a Cobb-Douglas form random utility function in expenditures on other goods and services and characteristics of the goods/services purchased.

$$U_{itj} = (y_{it} - p_{jt})^a G\ (x_{1jt}, x_{2jt}, \vartheta_{jt}, z_t, v_i) e^{\varepsilon_{ijt}} \tag{6}$$

Where G (.) is a function articulating $x_{1jt}, x_{2jt}, \vartheta_{jt}, z_t$ , $v_i$ with multiplications of exponentials. $y_{it}$ is the income, and $p_{jt}$ is the price of alternative j at time t, which is a part of the characteristics $x_{1jt}$ that consumers have different preferences.

[96] The outside good generally referes to the alternatives not in the sample. Based on Berry et al. (1995), the outside good is an aggregation of all other alternatives available in the market but not included in the sample. More specifically, in the context of this paper, it refers to all outside lenders of loans that are not in the sample. In the U.S., insurance companies (e.g. State Farm) can also make lendings in the form of loans to borrowers. Many small lenders (e.g. credit unions) that are not included in the sample are aggregated into the outside good.

By taking the logarithm of equation (6) and allowing for individual consumers interacting with both the inside good and the outside goods/services, the utility system becomes

$$u_{itj}=a * log\left(y_{it}-p_{jt}\right)+\bar{\beta} x_{1\,(\neq p) jt}+\ \gamma x_{2 jt}+\rho z_t+\vartheta_{jt}+\left[x_{1\,(\neq p) jt}\right]^{\prime} * \sum v_i+\varepsilon_{ijt}\ ,$$

$$u_{it0}=a * log\left(y_{it}\right)+\rho z_t+\vartheta_{0t}+\left[x_{10t}\right]^{\prime} * \sum v_{i0}+\varepsilon_{i0t}, \qquad (7)$$

As pointed out by Xu et al. (2016), customers won't necessarily have a unified preference on price. Therefore, by following Xu et al. (2016), I also allow interaction between price term and individual customers, which turns the utility system into

$$u_{itj}=a * log\left(y_{it}-p_{jt}\right)+\bar{\beta} x_{1jt}+\ \gamma x_{2jt}+\rho z_t+\vartheta_{jt}+\left[x_{1jt}\right]^{\prime} * \sum v_i+\varepsilon_{ijt}\ ,$$

$$u_{it0}=a * log\left(y_{it}\right)+\rho z_t+\vartheta_{0t}+\left[x_{10t}\right]^{\prime} * \sum v_{i0}\ +\varepsilon_{i0t}, \qquad (8)$$

According to Berry (1994), the reason why the utility function of the outside goods is different is that the characteristics of the outside goods that customers have the homogenous preferences on are unobserved and don't influence the estimation on the inside goods under the BLP framework.

System (8) denotes the baseline utility speciation used in the BLP estimation in this paper. As derived by Berry et al. (1995)., if the customers have unified preferences on the set of characteristics $x_{1jt}$, and hence there's an interaction between characteristics and individual customer, the utility system can be turned into a Logit model since all $v_i\ and\ v_{i0}$

become 0. Assuming $\varepsilon_{ijt}$ and $\varepsilon_{i0t}$ both follow Weibull Distribution function, since $[x_{10t}]' * \sum v_{i0}$ *and* $[x_{1jt}]' * \sum v_i$ both become 0, the whole system becomes

$$u_{itj} = a * log(y_{it} - p_{jt}) + \bar{\beta} x_{1jt} + \gamma x_{2jt} + \rho z_t + \vartheta_{jt} + \varepsilon_{ijt}$$

$$u_{it0} = a * log(y_{it}) + \rho z_t + \vartheta_{0t} + \varepsilon_{i0t} \tag{9}$$

The market share functions are

$$\text{S}\ (x_{1jt}, x_{2jt}, z_t, \vartheta_{jt}, v_i; \overline{\beta}, \gamma, \rho, \textstyle\sum) = \text{S}\ (x_{1jt}, x_{2jt}, z_t, \vartheta_{jt}; \overline{\beta}, \gamma, \rho) =$$

$$\frac{e^{\delta_{jt}}}{1+\sum_{J=1}^{J} e^{\delta_{jt}}} = \frac{e^{a*log(y_{it}-p_{jt})+\bar{\beta}x_{1jt}+\gamma x_{2jt}+\rho z_t+\vartheta_{jt}}}{1+\sum_{J=1}^{J} e^{a*log(y_{it}-p_{jt})+\bar{\beta}x_{1jt}+\gamma x_{2jt}+\rho z_t+\vartheta_{jt}}} \tag{10}$$

Based on McFadden (1973), specification (10) also implies

$$\delta_{jt} = \ln(s_{jt}) - \ln\ (s_{0t})$$

$$= a * log(y_{it} - p_{jt}) + \bar{\beta} x_{1jt} + \gamma x_{2jt} + \rho z_t + \vartheta_{jt} \tag{11}$$

Specification (11) can be directly estimated by using least square estimation procedures and are used as robustness checks in this paper.

When the interactions between the characteristics and the individual customers are allowed, $[x_{10t}]' * \sum v_{i0}$ *and* $[x_{1jt}]' * \sum v_i$ are no longer 0s, and the share functions become:

$$\text{S}\ (x_{1jt}, x_{2jt}, z_t, \vartheta_{jt}, v_i; \overline{\beta}, \gamma, \rho, \textstyle\sum\ ) = \frac{e^{\delta_{jt}+\mu_{jt}}}{1+\sum_{J=1}^{J} e^{\delta_{jt}+\mu_{jt}}}$$

$$=\frac{e^{a*log\left(y_{it}-p_{jt}\right)+\bar{\beta}x_{1jt}+\gamma x_{2jt}+\rho z_{t}+\vartheta_{jt}+\left[x_{1jt}\right]'*\Sigma v_{i}}}{1+\sum_{J=1}^{J} e^{a*log\left(y_{it}-p_{jt}\right)+\bar{\beta}x_{1jt}+\gamma x_{2jt}+\rho z_{t}+\vartheta_{jt}+\left[x_{1jt}\right]'*\Sigma v_{i}}} \quad (12)$$

The BPL estimation procedure is based on the share functions in the specification (12), which captures the reality that in the monopolists' competitive market, the market shares of banks are dependent on the heterogenous customer preferences as well as some homogenous preferences shared across different customer groups who obtain loans.

### *2.3.2 Hypotheses*

According literature regarding the research of preference of consumers over financial institutions mentioned above including Boyd et al. (1994), Zineldin (1996), Phuong Ta et al. (2000), Almossawi (2001), several characteristics of banks have impacts on the consumers choices over banks: 1, Reputation 2, Technology and convenience; 3, Services quality and efficiency; 4, Charges and interest costs (Price) [97]; 5, Loan availability. That implies customers seeking financial services will make make a utility maximization depending on their preferences over those characteristics. By putting all those factors into the discrete choice framework proposed by McFadden (1980) and Train (2003), all those factors will account for the choice of an individual. When the individual level choice is unobservable, the individual choice can be aggregated as the share of a specific product/service provided by a bank according to the structural estimation model proposed by Berry (1994). Then by using the log-transformation, as indicated by Berry (1994) and Berry (1995), the log-transformed share of the products/services is a linear function of all

[97] Generally, the charges and the interest costs are the expenses to the customers. Therefore, they can be interpreted as the price of a loan service offered by a bank.

the factors that determine the choices. By leveraging the estimation techniques proposed by Berry et al. (1995), Berry and Pakes (2007), and Nevo (2001), the price elasticity can be computed. By identifying the price elasticity, the impacts from interest rate changes can be simulated and evaluated.

Generally, the paper is concerned about whether the decision factors proposed by Boyd et al. (1994), Zineldin (1996), Phuong Ta et al. (2000) Almossawi (2001) are also the factors that determine the choices of the borrowers in the U.S. Therefore, the first set of hypotheses to test include

**H1 (a):** Reputation of the bank, measured by goodwill to total asset ratio, is a factor that positively influences the choice of borrowers and increases the market share of the bank.

**H1 (b):** Technology level of the bank and the convenience to customers, measured by the intangible asset per employee and number of employees, are factors that positively influence the choice of borrowers and increases the market share of the bank.

**H1 (c):** Services quality and efficiency, measured by average staff expenses and the number of employees, is a factor that positively influences the choice of borrowers and increases the market share of the bank.

**H1 (d):** Interest cost, measured by the average loan interest rate offered by the bank, is a factor that negatively influences the choice of borrowers and decreases the market share of the bank.

**H1 (e):** Loan availability, measured by total asset value and liquid assets to total asset ratio is a factor that positively influences the choice of borrowers and increases the market share of the bank.

Based on the heterogeneous preference assumption adopted by McFadden (1980), Train (2003), and Berry et al. (1995), the borrowers might have different preferences over the price factor. Therefore, the second hypothesis to test is

**H2:** Borrowers have different preferences over interest rates, and hence the impacts from interest rates on the demands of loans vary across individuals which implies random parameters on interest rate.

Based on the traditional microeconomics theory and industrial organization theory regarding the relationship between demand and price for a normal good[98], the price factor may cause changes in the demand. Taking this into the structural demand estimation framework, the changes in interest rates should cause changes in the shares of the banks in the loan market. That should reshape the market structure as well. Xu et al. (2016) show this phenomenon in their research about the medical service market. Therefore, the third hypothesis to test is

**H3:** The interest rate adjustments by the Federal Reserve Bank will lead to asymmetric[99] structural changes in the loan market.

---

[98] Details can be found in Tirole, Jean. *The theory of industrial organization*. MIT press, 1988. The demand is usually a function of price and when the price of a normal good increases, the demands will go down. When the price of a normal good decreases, the demands will go up.

[99] The word "asymmetric" here refers that the effects of an increase in interest rates on a bank's loan demand is different from the effects of a decrease in interest rates on a bank's loan demand in terms of magnitudes.

The price elasticity itself can measure the sensitivity of the bank when the price changes. In the sense of that, price elasticity can be another measurement of market power: If the entity has a higher price elasticity, it’s more easily influenced by the price competition and it’s more fragile, which means less market power. The earliest research conducted on the issue of the relationship between market power and risk preference of banks is done by Bresnahan (1982). In this most primitive research, the author uses the banks' data and concludes that firms in monopolistic markets are more risk- averse than firms in competitive markets, indicating a desire for the 'quiet life'. That implies the banks with higher market power should be risk-averse. In order to test whether this is true or not, I use the interest rate elasticity as a proxy for the market power to test the fourth hypothesis:

**H4:** Banks with lower own interest elasticities are more likely to be risk-averse.

Back to the end of the last century, Phillips (1995) finds out that the capital structure of firms is highly correlated with industry output. They also find that the price in the final good market will be significantly changed once the major firms in those industries change their financial leverages. That implies the financial structural decisions of firms should be correlated with their market power. To test whether this is true or not for the banks in the U.S., I test the following hypothesis:

**H5:** Banks with lower interest elasticities are more likely to have higher financial leverages.[100]

[100] Generally, the financial leverage refers to the debt equity ratio. I follow the general definition in this paper.

## 2.4 Data

### *2.4.1 Data sources and raw data*

The data that I am going to use is the yearly data obtained from the Bank Scope database, which is a private database that collects the data by combining the responses to the questionnaires from the banks and the annual report published by the banks. This comprehensive database is explicitly focusing on the data of banks and can provide precise firm-level information. My data set could be constructed as a two-dimension panel data, namely year and individual bank, of demand and service characteristics for different banks. The raw data contains 24997 banks over the 4-year period from 2013 to 2016. Given the fact that the general BLP model proposed by Berry et al. (1995) is static, the data during this period will be a good fit for the model because, during this period of time, the whole banking industry is going through steady growth and restoration.[101] That means there are no substantial structural changes during this period in the industry. If there is no discernible demand shock or income shock during that period, it can be safely inferred that there were no huge changes in the customers' preferences.[102] In that sense, the data of this time period should fit the settings of the BLP-type model well.

The raw data contains all specialties in the banking industry in the United States.[103] Since the focus of this paper is to investigate the demand for the loan, I only keep the banks

---

[101] From Figure 2, we can see that the ROA ratios and ROE ratios keep steady during that period, which means there shouldn't be huge supply shock or demand shock in that period for the industry.
[102] Based on the classic microeconomics theory on consumer preference and demand, i.e. Varian's Microeconomic Analysis (2016), chapter 6 & 7, when preference parameters remain the same, income factor remain the same, then demand should remain the same as well.
[103] Based on the data from Bankscope, the specialties in the U.S. of banks include investment banks, bank holding companies, commercial banks, credit unions and policy banks.

that have a specialty in commercial banking and exclude the banks/firms with other specialties. The commercial banks are the major financial institutions that give out consumer and commercial/industrial loans in the U.S. Based on the data provided by Bankscope, the total loan given out by commercial banks accounts for about 91% of the total loan value in the U.S.[104]

The data source of the loan market size information is obtained from the FDIC quarterly/annual report.[105] Typically, I follow the tradition of using the total book value of loans in the U.S. as the total market demand and I use the loan share as the market shares of the banks.

A data screening is performed on the raw dataset by following several criteria: 1, I dropped the banks that are no longer operating since 2012 because most of them have a huge portion of data missing 2, I dropped the observations that have missing data in total asset, number of employees, total value of domestic deposits (including time deposit, consumer deposit, and demand deposit), total value of domestic loans (including loans to government and consumers plus commercial/industrial loans), total revenue of interests on loans, total cost of interests on deposits, intangible asset, goodwill.[106] The missing data points don’t present a specific pattern. Therefore, they should be randomly missing.

After the data screening, I was left with 6289 bank-year observations with 1879 banks represented in the dataset. Due to the complexity of the BLP estimation procedure,

[104] Computed by using the book value of outstanding loan owned by the commercial banks divided by the total book value of outstanding loans owned by all banks regardless of specialty.
[105] The detailed information is attached in Appendix A Table 1.
[106] Generally, from the perspective of corporate finance, goodwill is a balance sheet item under the category of asset. It refers to the accumulation of intangible asset and capital due to the good reputation, fame and good enterprise images.

[107] It will be difficult or even impossible for a computer to do the simulation and contraction mapping for each observation, and then the optimization, especially with interpreted languages such as Matlab, R, or even Python. To overcome this computational problem, I picked the top 30 with the largest asset value and bottom 30 with the smallest asset value from each market and form a new sample. After the sample size is reduced to a manageable level, the computational difficulty gets resolved. Since I also control for the asset value as an exogenous variable in the specification, this shouldn't generate any selection bias. Besides, the BLP framework takes care of the unobserved choices well by introducing the outside good share, and hence, the banks that are not selected into the estimation sample are modelled as the outside goods and shouldn't affect the estimation results. As a robustness check, I also randomly selected 30 small banks.

### *2.4.2 Definition of a market*

As previously mentioned in section 2, in typical choice models, a market is defined as a "city-time" combination. However, since there is no available data showing which area(s) the banks are serving and from which region the customers of the banks are coming, there's no feasible way to model a geographic region as a market. Besides, here specifically in the loan market, there's a particular property of the banking industry: Generally, all banks are available to serve the customers across the country due to the usage of the Internet and other transmitting technology though many of them put a focus on

---

[107] Since the BLP estimation procedure is adopting a nested loop structure, the big-O of this estimation procedure is O ($n^2$). This is already the best scenario given by using matrix operations as much as possible. Even under this best scenario, it will end up with 6289*6289 computations. It will take 1.2542 years to finish one estimation. The big-O notation is a notation used to show the complexity (computing time and memory) of an algorithm. The previous research (e.g. David, 2006) using BLP framework contain no more than 1000 alternatives due to the consideration of computational complexity.

nearby areas.[108] Thus, I don't use the "city-time" combination as the market segmentation. Instead, I use the time period as the market segmentation in this paper, and I treat the whole country's financial need in one period (specifically a year in this paper) as the demand in each market.

This definition of the market is somewhat problematic since some small banks they only focus on the customers that are geographically nearby instead of trying to attract loan borrowers nationwide. At the same time, even the smallest bank in the sample have at least one branch outside of the state where its headquarters are located, suggesting that the market definition may not be unreasonable. [109]

### *2.4.3 Definition of market shares*

Based on the previous research regarding market structures of banks (e.g., Gilbert, 1984), there are several measurements of market shares of banks that are commonly used. In this paper, I calculate the market share by using the share of total loan book value in a specific market by taking the book value of outstanding loan of each bank divided by the total book value of the outstanding loans in that year across the whole U.S. Based on that previous literature, the banking industry follows a monopolist competition structure. Thus, I follow Xu et al. (2016) and define the loan product of each bank to be an alternative for customers.[110] I compute the market share for each bank by using its book value of loans

[108] Based on the information obtained from www.branchspot.com, even the smallest bank in the sample has at least one branches outside of the state where its headquarter is in, which implies the fact that nowadays the banks don't limit their operations in a specific area any more. Web services provided by loan lender search engines, such as www.lendinghome.com, make getting a loan from other regions/areas much easier.

[109] This implies, based on the information from www.branchspot.com, that the banks don't limit their operations in a specific area any more. Web services provided by loan lender search engines, such as www.lendinghome.com, make getting a loan from other regions/areas much easier.

[110] In Xu et.al (2016), banks are modeled as alternatives since the services provided by each bank are associated with the differentiated characteristics of the bank. Approximately, a service alternative can be represented by a bank.

divided by the total book value of loans in each market which is defined as a year in this paper. For instance, the book value of JP Morgan Chase in the year 2013 was 6.325 billion USD and the total book value of loans in the U.S. was 84.52 billion dollars. Then JP Morgan Chase's loan market share in the 2013 market was roughly 7%.

### *2.4.4 Variables of interests and indicators*

The variables of interest in this paper include the characteristics of banks which are the determinants of consumers' choices over banks. Based on the previous literature regarding the preference of consumers over financial institutions (Boyd et al, 1994, Zineldin, 1996, Phuong Ta et al, 2000, Almossawi, 2001), there are several characteristics of banks that have impacts on the consumers choices over banks: 1, Reputation; 2, Technology and convenience; 2, Services quality and efficiency; 3, Charges and interests costs; 4, Loan availability. The key independent variables which proxy for the loan service characteristics include: 1, The loan interest rates which is computed by using the total interest revenue on loans divided by the total book value of loans; 2, The service quality which can be indicated by the personnel expenses on each employee and asset value per employee since these reflects how well the employees are paid, how well the employees are trained and how many resources that each employee can leverage as indicated by Ho (2015); 3, the loan availability can be measured by the liquid asset to total asset ratio sine the liquid asset to total asset ratio reflects how much liquid asset the bank is holding and banks are lending money by giving out its liquid assets (in most cases cash); 4, The reputation can also proxied by the total asset value since it shows how much asset the bank has and how powerful the bank can be; 5, The reputation can also be indicated by good will to total asset ratio in the sense that goodwill reflects the value of a company's brand name, solid customer base, good customer relations, good employee relations, and any patents or proprietary technology.[111] If the ratio between goodwill to total asset is huge it means the firm's reputation carries a huge weight as an asset and it means the firm has good reputations, good technological level and good customer relations overall; 6, Aside from the goodwill, the technology level can be indicated by the intangible asset values per employee. Since the data don't contain the information of R&D expenses in its income statement and that is also not a standard item on the income statements of financial institutions, I can't calculate the R&D intensity, which has been commonly used as one of the technology level indicators. However, from the standing point of accounting, the R&D expenses will finally be converted into patents or copyright, which are usually included in the intangible assets; 7, the accessibility level can be indicated by the number of employees.

### *2.4.5 Summary statistics*

In the final sample, the banks/firms that have specialty other than commercial banks are excluded. The banks that have missing information in the variables of interest are also excluded from the final sample, as explained in the data screening section above. The final sample covers a time period of 4 years from 2013 to 2016. There are 1375 observations in the 2013 market, 1500 observations in the 2014 market, 1592 observations in the 2015 market and 1822 observations in the 2016 market. The summary statistics are shown in Table 7.

[111] Definition of item goodwill on balance sheets: https://www.investopedia.com/terms/g/goodwill.asp#ixzz53M2tScI1

**Table 7 – Summary statistics of the data[112]**

| | | Whole Sample | | | Subsample | | |
|---|---|---|---|---|---|---|---|
| Variable | Unit | N | Mean | Std. Dev. | N | Mean | Std. Dev. |
| Average Staff Expenses per employee | $ 1000 | 6,289 | 73.79 | 125.18 | 240 | 80.65 | 44.22 |
| Goodwill to total asset ratio | 1 | 6,289 | 0.02 | 0.03 | 240 | 0.03 | 0.03 |
| Average asset per employee | $1000 | 6,289 | 5714.02 | 14571.66 | 240 | 8784.3 | 27109.21 |
| Number of Employees | 1 | 6,289 | 1448.48 | 10880.12 | 240 | 22352 | 47156 |
| Average intangible asset per employee | $ 1000 | 6,289 | 0.02 | 0.10 | 240 | 0.03 | 0.03 |
| Total Asset | $ 1,000,000 | 6,289 | 10.50 | 927.00 | 240 | 181.00 | 406.00 |
| Liquid asset to total asset ratio | 1 | 6,289 | 7.93 | 9.11 | 240 | 12.97 | 12.07 |
| Loan interest rate | 1 | 6,289 | 0.05 | 0.02 | 240 | 0.05 | 0.02 |

As mentioned above, due to the computational difficulty, I have to select the banks with the largest assets and smallest assets to form a new sample to keep the sample size manageable. If we compare the summary statistics, the mean of most of the variables of interest in the two samples is close to each other except for the number of employees. The mean of the loan interest rates is almost completely the same between the two samples.

[112] Data sources: Bank Scope database. Computed by the author.

Besides, in the estimation, the size is controlled. Therefore, the smaller sample should be enough to identify the true parameters.

## 2.5 Empirical methodologies

To overcome the non-linear price endogeneity problem raised by Berry (1994), the details of my random coefficients-based structural model, including the setup, assumptions and identification strategies, will be elaborated in this section. More precisely, the estimation has to deal with three issues: (i) the lack of consumer individual-level choice and demographical information, (ii) the heterogeneity of tastes of customers, and (iii) the endogeneity problems caused by the correlation between service characteristics and unobserved banking demand shocks. I also capture the intense non-price competition in the banking industry that is caused by various characteristics.[113]

In terms of the first problem, the BLP framework was proposed to resolve the unobserved individual choices in the demand estimation problems: the BLP framework only requires the aggregated demand data for estimation. As for the issue of lacking in demographical information, based on the literature regarding the customers' choice over loan services mentioned in section 2.2 (e.g. Boyd et al., 1994), the most important determinants on choices of bank services are commonly shared by different types of consumers regardless of marital status, age, income or household members. Thus, it will be unnecessary to use the demographic distribution information to do the simulation. On the other hand, based upon the random utility framework, a different individual might have a different perspective on a specific characteristic, and I will do a simulation to relax the

[113] In this context, it refers to the competition caused by determinants other than interest rates offered and charged.

"same-taste" or so-called "same-preference" assumption as proposed by Nevo (2005). The elaboration on this approach is in section 2.5.1.

As for the endogenous problem, the endogeneity issue needs some particular identification strategies to resolve. The identification strategy with details is discussed in section 2.5.2.

### *2.5.1 Estimation techniques*

Based on the traditional BLP- type model, my goal is to estimate the mean utility function. In order to conduct the estimation, I still follow the classical method of Berry et al. (1995) and Nevo (2001). I use the generalized method of moments (GMM) estimation with optimized instrumental variables. The contraction mapping algorithm is strictly following the algorithm in Berry et al. (1995) and Nevo (2001).[114] The critical intuition and purpose behind the contraction mapping are 1, Ensure the convergence of optimization of the GMM estimation procedure; 2, Making sure the market share after the simulation of an alternative still matches the true market share. By making the share function become conditional only on characteristics, we have

$$\mathrm{S}\,(x_{1jt}, x_{2jt}, z_t, \vartheta_{jt}; \overline{\beta}, \gamma, \rho) = \int f\,(v_i,\ \delta_{jt}) P\,(dv) \tag{13}$$

where, $f\,(v_i, \delta_{jt})$ *is the joint distribution of* $v_i$ *and* $\delta_{jt}$ .

[114] Please note that the key difference between this model and BLP (1995) and Nevo (2001) is that their model is a dynamic one and Xu et al. (2016). Therefore, by following Xu et al. (2016), the simulation step is a combined with all markets and the different markets (over different years) are compiled into one when doing contraction mapping and simulations. This is specified in the notations.

As pointed out by Berry et al (1995)., specification (13) doesn't have a closed-form solution. In order to overcome this issue, based on the contraction mapping method proposed by Berry et al (1995)., the operator T which maps between share, parameters, and the simulations can be written as

$$T = \delta_{jt} + \ln\left(s_{jt}\right) - \ln\ \left(\mathrm{s}_j\left(x_{1jt}, x_{2jt}, z_t, \vartheta_{jt}; \overline{\beta}, \gamma, \rho\right)\right) \tag{14}$$

By the time of convergence of $\delta_{jt}$, combing with the simulation on the interaction term and assuming $v_i$ follow a specific (normal or uniform) distribution, the share conditional only on parameters characteristics can be written as

$$\mathrm{S}\ (x_{1jt}, x_{2jt}, z_t, \vartheta_{jt}; \overline{\beta}, \gamma, \rho) = \frac{1}{nst} * \sum_1^{nst} f\ (v_i\ , {\delta_{jt}}^*) \tag{15}$$

Where ${\delta_{jt}}^*$ denotes the convergence value of $\delta_{jt}$ after contraction mapping.

Then, by plugging in the shares in each iteration and the parameters in each iteration into the optional instrumental GMM objective function proposed in Berry et al. (1995) and then minimize the objective function will give the optimal GMM estimation of parameters $\overline{\beta}, \gamma, \rho$.[115]

### *2.5.2 Identification strategy*

Due to the existence of the unobserved characteristics, one of the key issues in the estimation is the endogeneity of the loan interest rates since the unobserved characteristics can influence the interest rate of loans and the share simultaneously. For

---

[115] The minimization method in Berry et al. (1995) uses the Nelder-Mead algorithm for minimum search. More details of this algorithm can be found in Berry et al. (1995) and Nevo (2001).

instance, the advertising and marketing costs can be the factors that determine the interest rates charged on loans and the market share simultaneously. However, those costs are not in the data. In order to resolve this issue, I introduce a type of instrumental variables (IVs). By following Berry et al. (1995), I use the Hausman-type IVs: the average of characteristics of other banks in the same market (T=t). Since the major operating costs of the banks are the interests paid to the people who make deposits and who lend money to the banks, by following Berry (1994), I also use the cost shifters of the banks: the deposit interest rates. To avoid the simultaneity issue, I use the average of the deposit interest rates of other banks in the same market (T=t). Though this Hausman-type IV has been widely used (e.g. Berry et al., 1995), there's still some notable restriction specifically for this dataset; The variation seems to be small across the IV vector, which might cause problems in estimation.

To sum up, the instrumental variables used in the estimation include 1, the average of the asset per employee of other banks in the same market, 2, the average of the goodwill to total asset ratio of other banks in the same market, 3, the intangible asset of other banks in the same market, 4, the average of staff expenses per employee of other banks in the same market and the average of the interest rates on deposits of other banks in the same market, and 5, the average of the deposit interest rate of other banks in the same market.[116]

The underlying identification assumption is that the characteristic of suppliers in the same market will have impacts on the pricing strategy of each other, while those characteristics will not have direct impacts on the market share of each other. Also, I assume the characteristics of other banks will not directly influence the share of a specific

[116] The deposit interest rate of each bank is computed by using the total deposit interest expenses on the income statement divided by the total book value of deposit on the balance sheet of that year.

bank in the current period, but they will have some impacts on the pricing strategy of that bank in the current period. These assumptions are mostly following the rationale of Berry et al. (1995). Thereby, the characteristics of other banks in the same market will be correlated with the loan interest rates and service charges but will not be correlated with the market share directly. If the discrete choice framework and the BLP-type choice model correctly captures the underlying data generating process, the estimators should indicate the actual marginal effect from banks' characteristics on market share.

## 2.6 Empirical results and implications

### *2.6.1 Results from the basic model*

In this subsection, the structural estimation results are given as well as the reduced form regression. The reduced form estimator should serve as a robustness check since it's simply assuming in the random utility of each customer, the preferences of customers over the price factor are homogeneous. The reduced form estimator imposes an additional assumption that all customers have the same preferences over the bank characteristics. This assumption may induce some slight biases but should not influence the significance too much if the model did capture the underlying structure of the data generating process.[117] Thus, the reduced form model should serve as a robustness check for the significance of the estimators obtained from the structural estimation are significant factors that determine the choice of borrowers.

[117] Based on the rationale of asymptotic theory, all unbiased estimators should converge to the same value as indicated by the law of large numbers. Andersen (1970) proves the asymptotic property of the MLE estimators, which include the logit model estimator, and hence the BLP estimators should also has this asymptotic property.

Table 8 shows the results of the BLP estimation results and Table 9 shows the two-stage least square results.

**Table 8 – BLP estimation results[118]**

| | ***BLP Estimation*** | | | |
|---|---|---|---|---|
| | *Dependent variable:* | | | |
| | $\log(share_{it}) - \log(share_{0t})$ (1) | | $\log(share_{it}) - \log(share_{0t})$ (2) | |
| | ***BLP estimation*** | $\Sigma$ | ***BLP estimation*** | $\Sigma$ |
| Loan interest rate | **-73.55*** (20.86)** | 0.32 (361.49) | **-75.32** (34.08)** | 6.75 (46.11) |
| Number of employees | 0.000063*** (0.000018) | | 0.000041*** (5.19e-06) | |
| Goodwill to total asset ratio | 20.67 (172.99) | 0.88 (504.48) | | |
| Intangible asset per employee | 21.37*** (8.03) | 7.89e-06 (367.9399) | 24.27** (11.95) | 0.08 (2005.00) |
| Liquid asset to total asset ratio | -0.063*** (0.019) | | -0.07*** (0.02) | |
| Asset per employee | -8.86e-07 (0.000014) | | | |
| Average staff expenses | 0.03*** (0.01) | | 0.02*** (0.01) | |
| Total asset | -0.273 (0.21) | | | |
| constant | -6.98*** (1.58) | | -5.91*** (1.78) | |
| N | 240 | | 240 | |
| Markets | 4 | | 4 | |
| Halton draws | 1000 | | 1000 | |
| f (p) ojective function value | 110.23455 | | 163.1526 | |

118 The numbers in the parentheses are GMM standard errors. The estimation procedure strictly follows the simulation and GMM estimation procedure proposed by Berry et al. (1995). $\Sigma$ $d$enotes the variance of the coefficient. The numbers in the parenthesis are the standard error of the estimators. The whole sample estimation means that the estimation is based on the 6289 observations without any selection. The subsample means that the estimation is based on the sample which contains the 30 largest and the 30 smallest banks.
Data sources: Bank Scope database. Computed by the author.

**Table 9 – 2SLS estimation results[119]**

***Subsample v.s. Whole Sample***

| | Dependent variable: $\log(share_{it}) - \log(share_{0t})$ | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| | ***2SLS estimation with sub- sample (selected by size)*** | ***2SLS estimation with sub- sample (selected by size)*** | ***2SLS estimation with whole sample*** |
| Loan interest rate | **-73.55***** | **-72.73***** | **-103.54***** |
| | **(20.87)** | **(18.24)** | **(17.07)** |
| Number of employees | 0.000066*** | 0.000041*** | 0.0001*** |
| | (0.00002) | (4.45e-06) | (0.00003) |
| Goodwill to total asset ratio | 26.49** | 25.79** | 6.19*** |
| | (10.47) | (10.53) | (1.55) |
| Intangible asset per employee | 20.78*** | 23.02*** | 0.15 |
| | (7.26) | (6.56) | (0.70) |
| liquid asset to total asset raio | -0.06*** | -0.07*** | -0.01 |
| | (0.02) | (0.01) | (0.008) |
| Asset per employee | -1.91e-06 | | 0.0001** |
| | (0.00002) | | (4.84e-06 ) |

[119] The numbers in the parentheses are the 2SLS standard errors. The model estimated on is the reduced form specification proposed by Berry et al. (1995). The key difference between this estimation and the BLP estimation is that the 2SLS estimation doesn't assume the customers have different preferences over the product characteristics. The whole sample estimation means that the estimation is based on the 6289 observations without any selection. The subsample means that the estimation is based on the sample which contains the 30 largest and the 30 smallest banks.
Data sources: Bank Scope database. Computed by the author.

**Table 9 – 2SLS estimation results (continued)**

| | | | |
|---|---|---|---|
| Average staff expenses | 0.03** | 0.03*** | -0.0003 |
| | (0.01) | (0.006) | (0.0006) |
| Total asset | -0.273 | | -0.99*** |
| | (0.21) | | (0.32) |
| constant | -8.06*** | -7.87*** | -4.42*** |
| | (1.84) | (1.22) | (0.86) |
| **1st stage** | | | |
| IV1-deposit interest rate | -67.58*** | -67.63*** | 2.98*** |
| | (8.58) | (8.56) | (1.00) |
| IV2-staff expensese per employee | -0.009*** | -0.006** | 0.00002 |
| | (0.002) | (0.003) | (0.00005) |
| IV3- goodwill | 7.02*** | 6.48*** | 6.44*** |
| | (1.12) | (1.34) | (0.30) |
| IV4- asset per employee | -4.31e-06* | -9.67e-06** | -2.90e-08 |
| | (2.28e-06) | (3.81e-06) | (5.97e-07) |
| IV5- intagible asset per employee | 47.69*** | 3.55 | -43.61*** |
| | (7.60) | (2.84) | (14.94) |
| N | 240 | 240 | 6289 |
| F (8,N-9) | 99999 | 47.29 | 265.94 |
| 1st stage R-sq | 0.3925 | 0.3659 | 0.3474 |
| R-sq | 0.4659 | 0.4625 | 0.2303 |

As presented by Table 8 and Table 9, the loan customers of the banks do consider the interest rate of loans, as expected. Additionally, they consider the accessibility, convenience and service quality: the estimated coefficients of the number of employees, intangible asset per employee and average staff expenses are all positively significant in both models. Most of them are significant at the confidence level of 99%. However, the estimated coefficient of the term goodwill to asset ratio is not significant in the BLP estimation model but significant in the multinomial-logit model which is estimated by 2SLS though the magnitudes of the estimators of both models are fairly close to each other. Surprisingly, the estimated coefficients of liquid assets to a total asset in all models are negatively significant, which means the loan borrowers are more likely to choose the banks with less sufficient funds available for loans. This finding is, to a certain extent, counterintuitive, and the reasons for this are not clear. A possible interpretation of this can be that there are omitted variable(s) (e.g. minimum deposits, other bank requirements) positively correlated with *liqassets/totassets* but negatively correlated with choices that biased the estimators. This finding may also suggest that smaller banks may be preferable on some margins like the quality of service, despite their sizes.

As expected, the sign of the coefficient of the loan interest rate is negative. The magnitudes of all models are fairly close to each other if we look at the estimation results by using the small samples. The coefficient ranges from -72 to -75. Essentially, this number means on average, a bank will experience a decrease of approximately 0.72% to 0.75% in the odds of its original share in the loan market relative to the share of the outside lenders if its loan interest rate increases by 0.01 in magnitude. The sign of the number of employees, intangible assets per employee, and average staff expenses are all positive as expected. The magnitude of the coefficient of term intangible asset per employee is unexpectedly large, which implies that customers do consider the technology and the degree of service convenience a lot. Based on the coefficients of this term in the BLP model, we can anticipate that, on average, a bank will gain around a 21% increase in the odds of its original share relative to the share of outside lenders if its average intangible asset per employee goes up by $1000.  If we look at the service quality side, we can see that the loan customers also consider the service quality a lot. The magnitude of the coefficient of the term staff expenses per employ is around 0.026, which implies that on average, the odds between the loan market share of a bank relative to the share of outside lenders will go up by 0.026% if the average staff expenses go up by $1000 dollars. To sum up, the reputation of the bank, the technology level of the bank and the convenience to customers, services quality and efficiency, charges and interest costs are significant factors that determine the choice of the borrower while the loan availability is not.

It's a little surprising that the amount of assets that a bank has doesn't have impacts on the share of that bank in the loan market. This result is different from the findings in Ho (2015) regarding the Chinese deposit market. However, one interesting finding is that when the asset term interacts with the price term, the interaction term becomes significant. That is to say, the asset values/size itself is not a determinant of demands but it becomes a determinant through price effects.[120] Though the findings are different, they are not necessarily contradictory in the sense that Ho (2015) is looking at the deposit market and I

[120] Details can be found in the robustness check section 2.7.2.

am looking at the loan market. Besides, the difference between the preference of Chinese customers can be quite different from the preference of U.S. customers.

There's one thing that is worth attention in the estimation results: The standard deviations of the characteristics variables are not significant and the magnitude of them is fairly small. That is saying, the opinions and perceptions of customers on those characteristics are not quite different and hence, the random coefficient logit utility model should be the same with a simple multinomial logit utility model essentially. That is to say, H2 is rejected.

Based on Berry (1994), since the standard deviation estimators in the random coefficient utility model is not significant, then the coefficients can be safely estimated by simply using the 2SLS estimation procedure and treat the choice and utility from the choices of consumers as multinomial logit utility system. Then, I also perform another estimation for the complete sample. As shown in Table 2, many coefficients are very close to the corresponsive ones in the BLP estimation. The coefficients of loan interest rates are slightly different and the whole sample gives a number with an even higher magnitude. However, the estimator from the whole sample is still in the 95% confidence interval of the BLP estimator. Therefore, we shouldn't think the estimation results from the smaller sample is biased by the sample selection procedure.

### *2.6.2 Counterfactual experiments with monetary policies*

The estimation results above make scenario analysis possible providing the loan interest rate elasticity. It's possible to test how the market share will change if the interest rate changes. By calculating the BLP elasticity for loan interest rates, predictions for shares changes of all banks in the industry will be doable. For the elasticity computation, we can either use the BLP elasticity or the standard logit elasticity based on the baseline results in Table 8 and Table 9. In this section, we use the BLP elasticity in these counterfactual experiments. For the sake of simplicity, I only choose the 30 largest banks that have the largest share and 30 smallest banks that have the smallest share in the sample to explore the effects of interest rate shocks and to show the patterns. To obtain the most up-to-date information, this section only focuses on the 2016 market.

The computation of the cross-elasticities and own-elasticities are following the method proposed by Berry et al. (1995).

$$\mathrm{e}_{jkt} = \begin{cases} -\frac{p_{jt}}{S_{jt}} \int a_i Pr_{ijt}\left(1 - Pr_{ijt}\right) dF(v_i) \ \ if\ j = k \\ \quad \frac{p_{kt}}{S_{jt}} \int a_i Pr_{ijt} Pr_{ikt} dF(v_i) \ \ if\ j \neq k \end{cases}$$

Where $p_{jt}$ denotes the interest rate on loans of bank j in market t in this paper. $S_{jt}$ is the actual loan market share of bank j at time t. $a_i$ is the coefficient of the loan interest rate estimated for individual i based on the simulation. $Pr_{ijt}$ is the probability of choosing j by individual i at time t according to simulation. $F(v_i)$ is the distribution of the deviation from mean utility $v_i$.

In order to explore the impacts of the monetary policies on the structure of the loan market in the U.S., I only display the 5 biggest banks and 5 smallest banks here in the elasticity table. The elasticities computed are shown below in Table 10. In Table 10, the cross and own-interest rate elasticities of the largest 5 banks and the smallest 5 banks are presented.

## Table 10 – Cross and Own-interest Rate Elasticities[121]

***Cross and Own-interest Rate Elasticities***

| | JPMorgan Chase Bank, NA | Wells Fargo Bank, NA | Bank of America, National Association | Citibank NA | Wachovia Bank, National Association | The Harvard State Bank | The Reedsburg Bank | The First National Bank of Athens | Teton Banks | FirstBank of Nebrska |
|---|---|---|---|---|---|---|---|---|---|---|
| JPMorgan Chase Bank, NA | -2.22018 | 0.350886 | 0.312454 | 0.329714 | 0.138106 | 5.83E-05 | 0.000114 | 5.4E-05 | 3.98E-05 | 6.07E-05 |
| Wells Fargo Bank, NA | 0.243653 | -2.62847 | 0.312454 | 0.329715 | 0.138106 | 5.83E-05 | 0.000114 | 5.4E-05 | 3.98E-05 | 6.07E-05 |
| Bank of America, National Association | 0.243649 | 0.350881 | -2.50317 | 0.329706 | 0.138105 | 5.83E-05 | 0.000114 | 5.4E-05 | 3.98E-05 | 6.07E-05 |
| Citibank NA | 0.243653 | 0.350887 | 0.31245 | -4.11448 | 0.138105 | 5.83E-05 | 0.000114 | 5.4E-05 | 3.98E-05 | 6.07E-05 |
| Wachovia Bank, National Association | 0.243652 | 0.350885 | 0.312456 | 0.329713 | -3.15588 | 5.83E-05 | 0.000114 | 5.4E-05 | 3.98E-05 | 6.07E-05 |
| The Harvard State Bank | 0.243653 | 0.350886 | 0.312452 | 0.329717 | 0.138105 | -3.72196 | 0.000114 | 5.4E-05 | 3.98E-05 | 6.07E-05 |
| The Reedsburg Bank | 0.24364 | 0.350869 | 0.312484 | 0.32969 | 0.138105 | 5.83E-05 | -5.61885 | 5.4E-05 | 3.98E-05 | 6.07E-05 |
| The First National Bank of Athens | 0.243653 | 0.350887 | 0.312449 | 0.329719 | 0.138105 | 5.83E-05 | 0.000114 | -4.46533 | 3.98E-05 | 6.07E-05 |
| Teton Banks | 0.243651 | 0.350883 | 0.31246 | 0.32971 | 0.138105 | 5.83E-05 | 0.000114 | 5.4E-05 | -2.83525 | 6.07E-05 |
| FirstBank of Nebrska | 0.243653 | 0.350887 | 0.312449 | 0.329719 | 0.138105 | 5.83E-05 | 0.000114 | 5.4E-05 | 3.98E-05 | -4.0752 |

As indicated by the table, by looking at the own-loan interest rate elasticities, the larger banks are less sensitive to the loan interest rate changes. For instance, JP Morgan Chase, the largest bank in the U.S. has an own loan interest rate elasticity is much lower than that of the smallest bank in the sample, FirstBank of Nebraska. Generally, larger banks are much less sensitive to loan interest rate changes. If we go back and look at the sample for the BLP regression, the 30 biggest banks (by asset) have an average of own-interest rate elasticities of -3.20 and the 30 smallest banks (by asset) have an average of own-interest rate elasticities of -3.89. If we only look at the top 5 and the bottom 5, the difference is even larger: The average of own-interest rate elasticities of the largest five

[121] Table 10 shows the cross and own-interest rate elasticities among the subsample which consists of the largest 5 banks and smallest 5 banks in terms of total asset value. The elasticities are computed with simulations of random effects. The dot line separates the results by size. This computation is based on the BLP estimation and simulation results with the 30 largest and 30 smallest banks. Sources: Computed by the author.

banks is -2.92, and the average of own-interest rate elasticities of smallest five banks is -4.14. The own-interest rate elasticities of the largest 30 banks and the smallest 30 banks are presented in Figure 4. The boxplots of the own-interest rate elasticities of the largest 30 banks and the smallest 30 by years are shown in Figure 8.

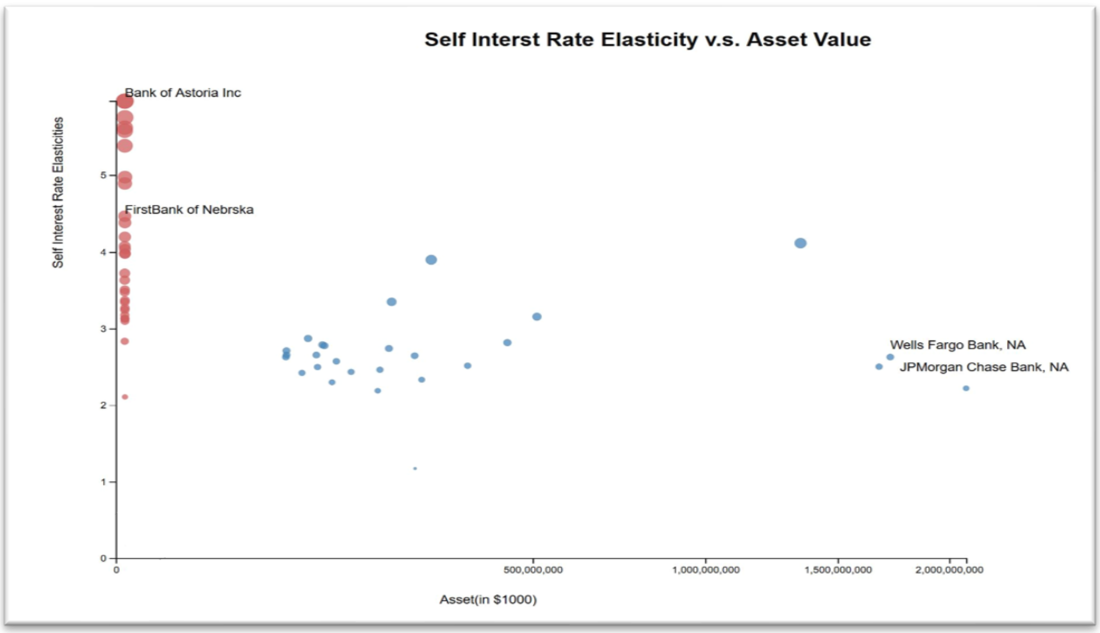


**Figure 8 – Own interest rate elasticity.**[122]

[122] Notes: 1, The large group, denoted by blue dots, consists of the 30 largest banks of year 2016 in sample, and the small group, denoted by red dots, consists of the 30 smallest banks of year 2016 in the sample. The size of the bank is measured by total asset value. 2, The elasticity presented in the graphs is the absolute value. 3, The bubble size is weighted by the elasticity. Sources: Computed by the author.

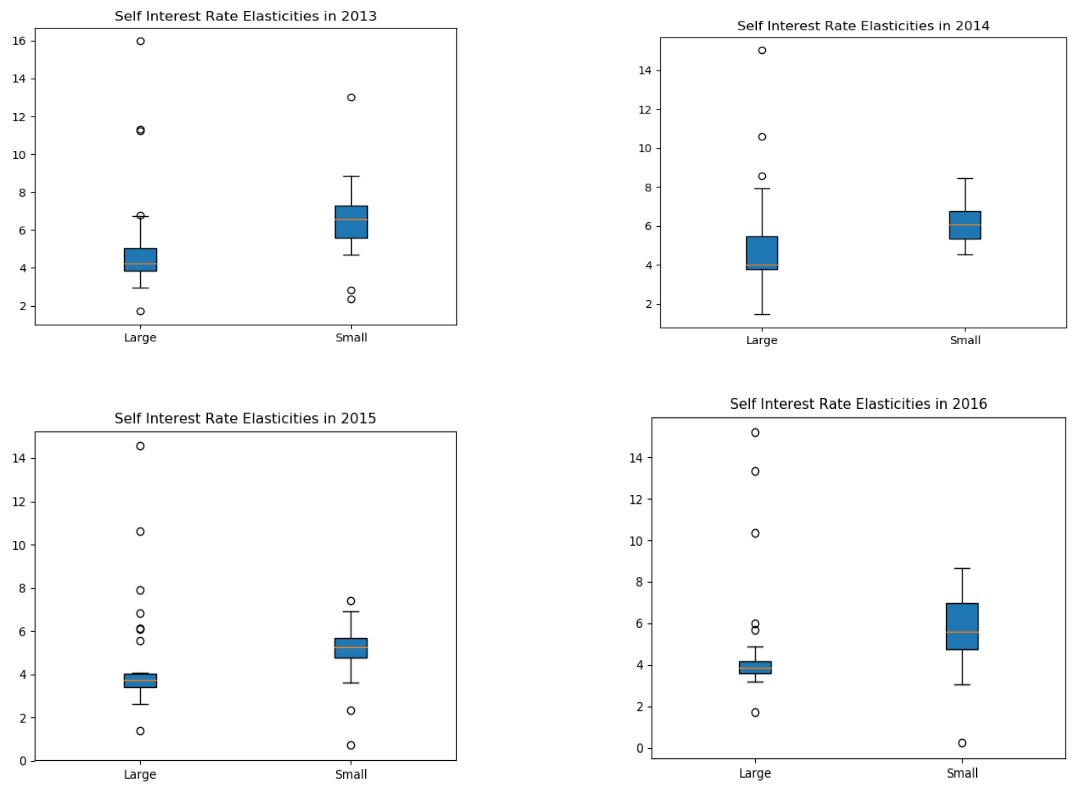


**Figure 9 –** ***Own interest rate elasticity over time boxplots***

Figure 8 and Figure 9 indicate that small banks are generally more sensitive to interest rates and small banks can lose more market share in the loan market than the large banks if the central bank, namely the FED, changes the base interest rate.

As for the outside good elasticity, I also compute the cross elasticities of how the interest rate of loans of outside lenders can change the loan market share of the banks in the sample. By following Berry et al. (1995) and Nevo (2001), the elasticities of the outside goods are computed by doing simulations on $e_{j0t} = -\frac{p_{0t}}{s_{jt}} \int a_i Pr_{ijt} S_{0t} dF(v_i)$ where $p_{0t}$ is computed by using the average loan interest rates of all the banks in the U.S. at time t. The cross-interest rate elasticities for the biggest 5 and smallest 5 banks with outside lenders are shown in Table 11.

## Table 11 – Scenario Analysis [123]

*Panel (a) Cross interest rate elasticity with outside lender*

| Bankname | Cross interest rate elasticities with outside lenders | Market gain from outside lender if the interest rate of loans from outside lenders goes up by 0.01 |
|---|---|---|
| JPMorgan Chase Bank, NA | 0.55959 | 0.00560 |
| Wells Fargo Bank, NA | 0.67667 | 0.00677 |
| Bank of America, National Association | 0.63949 | 0.00639 |
| Citibank NA | 1.00937 | 0.01009 |
| Wachovia Bank, National Association | 0.82495 | 0.00825 |
| The Harvard State Bank | 0.84535 | 0.00845 |
| The Reedsburg Bank | 1.27619 | 0.01276 |
| The First National Bank of Athens | 1.01419 | 0.01014 |
| Teton Banks | 0.64395 | 0.00644 |
| FirstBank of Nebrska | 0.92558 | 0.00926 |

[123] Table 11 panel (a) shows the cross-elasticity of the 5 largest banks and the 5 smallest banks with the outside lenders. That panel also indicates how those banks obtain market shares within the loan market if the outside lenders increase the interest rates. Table 11 panel (b) and panel (c) shows the scenario analysis about how the market shares of the largest banks and the smallest banks can be influenced by the Federal Reserve's decision of adjusting interest rates. The calculation is based upon the own-elasticity, the cross-elasticity by all other firms and the cross-elasticity by outside lenders. The dot line separates the large banks and small banks. Sources: Computed by the author.

## Table 11 – Scenario Analysis (continued)

***Panel (b) Market Share Changes after a 0.01 (100 base points) Increase on Loan Interest Rates***

| Bankname | Market share loss if the interest rate of loans goes up by 0.01 nationally (%) |
|---|---|
| JPMorgan Chase Bank, NA | -0.00529 |
| Wells Fargo Bank, NA | -0.00928 |
| Bank of America, National Association | -0.00801 |
| Citibank NA | -0.02060 |
| Wachovia Bank, National Association | -0.01094 |
| The Harvard State Bank | -0.01502 |
| The Reedsburg Bank | -0.02968 |
| The First National Bank of Athens | -0.02076 |
| Teton Banks | -0.00816 |
| FirstBank of Nebrska | -0.01775 |

***Panel (c) Market Share Changes after a 0.005 (50 base points) Decrease on Loan Interest Rates***

| Bankname | Market share loss if the interest rate of loans goes down by 0.005 nationally (%) |
|---|---|
| JPMorgan Chase Bank, NA | 0.00265 |
| Wells Fargo Bank, NA | 0.00464 |
| Bank of America, National Association | 0.00401 |
| Citibank NA | 0.01030 |
| Wachovia Bank, National Association | 0.00585 |
| The Harvard State Bank | 0.00751 |
| The Reedsburg Bank | 0.01484 |
| The First National Bank of Athens | 0.01038 |
| Teton Banks | 0.00408 |
| FirstBank of Nebrska | 0.00887 |

Based on Table 11 panel (a), the largest banks will gain more market shares if the outside lenders increase their loan interest rates by 0.01 than the smallest banks do. This finding shows that small banks are more sensitive to the interest rate changes by the outside lenders.

In order to elaborate on the effects of changes in interest rate imposed by monetary policies, I simulate three different scenarios here:

Scenario 1: A monetary policy sets the basic loan interest rate 100 basis points higher

In this scenario analysis, I only pick out the largest and smallest 5 banks in the sample, and I especially focus on JPMorgan Chase Bank and FirstBank of Nebraska to illustrate the impacts. If all banks follow this monetary policy by increasing 1% in loan interest rates for all loans issued and holding everything else constant, then by calculating the cross-loan interest rates and own loan interest rate elasticities, market share changes can be computed under that monetary policy.

From the computed results presented in Table 11 penal (b), JP Morgan Chase will lose 0.0053% of its share in the loan market while the smallest bank in the 2016 sample FirstBank of Nebraska will lose 0.01775% of its market share. This is saying, the monetary policy that raises the loan interest rate will make the largest bank lose less market share while making the smallest banks losing more customers and market share. That is to say, the monetary policy of increasing interest rates may lead to a more unbalanced market structure for the loan market and even the commercial banking industry. It will squeeze out the smallest banks and have trivial impacts on the largest banks.

Scenario 2: A monetary policy sets the basic interest rate 50 basis points lower

In this scenario analysis, I only pick out the largest and smallest 5 banks in the sample, and I especially focus on JPMorgan Chase Bank and FirstBank of Nebraska to illustrate the impacts. If all banks follow this monetary policy by decreasing 0.5% in loan interest rates for all loans issued and holding everything else constant, then by calculating the cross loan interest rates and own loan interest rate elasticities, we can see the market share changes under that monetary policy in Table 11 panel (c).

Interestingly, from the calculation results, we can see that JP Morgan Chase will be gaining 0.0026% more market share while the smallest bank FirstBank of Nebraska will be gaining 0.00887%, which is 4 times the increase of JP Morgan Chase will experience. By doing this calculation, we can see that monetary policy that decreases the loan interest rate even a little bit, will give the small banks a decent growth in the market share. If the interest rate decrease is large, 200 base points (2%) for instance, then the smallest banks will gain more market share in the loan market.

Up to this point, based on the experiment results presented above, we fail to reject H3. Based on these experiments, when the Federal Reserve Bank changes the base interest rate, the smallest banks might face survival challenges due to the "squeezing effects" from the large banks, causing the market structure to change.

## 2.7 Robustness check

### *2.7.1 BLP framework with the control function approach*

As noted in the identification strategy section, the BLP framework is using instrumental variables combined with GMM to implement the estimation procedure. Based on Berry et al. (1995) and Nevo (2001), the reason for using the IV-GMM estimation is that the endogeneity entered in a non-linear way. To identify the coefficient consistently, the IV-GMM method requires that there exists a linear relationship between the IV, and that the endogenous variable, and that there also exists a linear relationship between the IV and the dependent variable as well. However, the control function approach can relax these requirements and gives more consistent estimators when the linear assumptions about the IV do not hold. Following the procedure proposed by Petrin and Train (2010), since the BLP estimation with the IV-GMM method has proved the homogeneity of the preference over the price factor, namely the loan interest rate, the control function approach here follows the standard control function approach.

Due to the homogeneity of customers' preferences over the price factor implied by the estimation on the BLP model, there's no need to include the random effect term in the model. By imposing the standard assumption of Control function approach:

1, $Price_{it} = \boldsymbol{Z}'_{\boldsymbol{it}}\theta + v_{it}$

2, $E\left[\varepsilon_{ijt} \middle| Price_{it}\right] = E\left[\varepsilon_{ijt} \middle| v_{it}\right] = \omega v_{it}$

estimations will be conducted on the following equation:

$$y_{it} = a + \beta_1 * Price_{it} + \boldsymbol{X_{it}}'\vartheta + \omega v_{it} + e_{it}$$

Where $y_{it}$ denotes the dependent variable which is computed through $\log(share_{it}) - \log(share_{0t})$ based on the BLP structure, $\mathrm{Price}_{it}$ denotes the price factor which is the loan interest rate, $\boldsymbol{X_{it}}$ denotes the vectors of other determinants that might influence the market share, $v_{it}$ denotes the control of the endogeneity of the price[124], and $e_{it}$ is the new random error term.

The estimation results are shown in Table 1 in Appendix B. As indicated, the coefficient of the price factor obtained from the control function approach estimation is not significantly different from the 2SLS estimation results and the BLP estimation results: The magnitudes are quite similar. Therefore, the estimators of the interest rate elasticity should be robust.

### *2.7.2 Interaction term with size and loan interest rate*

As a part of the key findings and contributions of this paper, I show that the impacts from changes on interest rate have different effects on banks with different sizes. In order to ensure the robustness of this finding, I also test the effects of the interaction term between the size and loan interest rate.

The estimations will be conducted on the following equation:

$$y_{it} = a + \beta_1 * Price_{it} + \boldsymbol{X}'_{\boldsymbol{it}}\vartheta + \beta_2 * Price_{it} * Size_{it} + e_{it}$$

[124] Mathematical details regarding the control function implementation procedure can be found in Woodridge's NBER notes, which can be obtained from https://www.nber.org/WNE/lect_6_controlfuncs.pdf

The results are reported in Table 2 of Appendix B.

As indicated in Table 2 of Appendix B, I tried two types of interaction terms. The first type of interaction is to interact with the loan interest rate with the asset value. The other type of interaction is to interact the loan interest rate with the size dummy. I generate the size dummy by comparing the assets of the bank to the mean value of assets: If the asset value of the bank is higher than the average asset value at time t, the size dummy equals to 1. Otherwise, it equals to 0.

Demonstrated by Table 2 of Appendix B, the interaction terms are always statistically significant. Besides, the estimations confirm our conclusion: when the asset values get larger or the size dummy equals to 1, the interest rate elasticities will get smaller. This aligns with the baseline results regarding interest rate elasticities in this paper.

### *2.7.3 Random sampling the small bank*

In the data section, I discussed that due to the computational complexity of the BLP algorithm, I had to select the largest 30 and smallest 30 banks to build up a smaller subsample that the computer can handle in the estimation. Even though the BLP algorithm will treat the observations not selected as outside goods, and that controlling the size in the estimation should address the selection bias issue, in order to ensure the robustness of the results and the conclusions, I also random sample the banks that fall into the bucket of small banks.

More precisely, I did the following:

1, Determine the quintile cut-off of the size, which is measured by the total book value of assets;

2, Keep the banks that fall into the 0-25% quintile in each year;

3, Use a uniform distribution to sample 30 banks from the list of banks obtained from step 2 randomly;

4, Merge the records of the banks selected from step 3 with the 30 largest banks of each year;

5, Conduct a BLP estimation on the sample obtained from step 4.

The reason why I only randomly sample the small banks is that the data show the lists of the largest banks, measured by using the total value of assets, don't change much from year to year since the number of the large banks is relatively small. Therefore, the 30 largest banks selected in the subsample used to generate the baseline results should be enough to represent the population of the large banks. However, since there are so many small banks in the United States, the 30 smallest banks might not be enough to represent the small banks. The BLP estimation results are attached in Table 3(a) in Appendix B.

As indicated by the results in Table 3(a) in Appendix B, the estimators are not significantly different from the baseline results.[125] The elasticities computed based on the estimators in Table 3(a) in Appendix B are still indicating that the smaller banks in this

[125] Since the BLP algorithm decomposed the variance-covariance matrix, there's no feasible way to test whether the two estimators are statistically different. However, if we look at the numbers in Table 10 and Table 2, the coefficient of the loan interest rate term in Table 2 is less than 2 standard deviations away from the estimator of the coefficient of the loan interest rate term in Table 10. Therefore, they should not be seen as statistically different.

randomly selected sample are more sensitive to interest rate adjustments than the large banks, which confirms the robustness of the conclusions.

Besides, in order to confirm the conclusion that the large banks are less sensitive than the small banks in terms of interest rates, I also conduct a t-test on the own-interest rate elasticities computed on this sample. The results are attached in Table 3(b) in Appendix B.

From the test results, it can be concluded that large banks have significantly lower own-interest rate elasticities than the small banks. This conclusion still holds when using random samples.

## 2.8 Market power and firm strategies

### *2.8.1 Market power and risk preference*

There exists a few previous research showing that the market power of a firm is highly correlated with managerial slack. The earliest research conducted on the issue of the relationship between market power and risk preference of banks is done by Bresnahan (1982). In this research, the author uses the banks' data and concludes that firms in monopolistic markets are more risk-averse than firms in competitive markets, indicating a desire for the 'quiet life'. That implies the banks with higher market power are more risk-averse. The most influential paper is Keeley (1990). In Keeley (1990), the author finds out that the market power of a bank is highly correlated with its risk preference in the sense that the banks who have more market power, as reflected in larger market-to-book asset ratios, hold more capital relative to assets (on a market-value basis) and they have a lower default risk as reflected in lower risk premiums on large, uninsured CD's. Another paper that tries to explore how risk preference is correlated with market power is Salas and Saurina (2003). They find that lower market power which is indicated by low economic profits reduces the incentives of banks to own--restrain from taking risks raises important regulatory issues in a period of liberalization and expectations of increases in competition among European banks.

In order to test if the American banks also follow the rule that banks with higher market power also are more risk-averse and prefer a quiet life, I use the empirical framework developed by Salas and Saurina (2003). I use the elasticity of loan interest rates as the measurement of market power. The measurement of risk exposure is borrowed from Salas and Saurina (2003) and Keeley (1990). In addition to the loan loss ratio used by Salas and Saurina (2003), I also looked at the tier 1 asset ratio and the subordinated debt asset to total asset ratio as the measurement of risks.

For the estimation in this part, since we have confirmed that the preference of the customers over the interest rate is not drastically different and the deviation from the mean utility is not significant, we can compute the own elasticities for all banks in the complete sample by using the simple multinomial logit elasticity without getting simulations involved. The elasticity calculation is based on Berry (1994).

More specifically, an estimation of the following equation is conducted:

$$y_{it} = a + \beta * MP_{it} + \boldsymbol{X'}_{\boldsymbol{it}}\gamma + e_{it} \tag{7}$$

Where $y_{it}$ denotes the dependent variables which measure the risk preference of the bank, $MP_{it}$ denotes the market power which is measured either by HHI or Own- Interest rate elasticity, $\boldsymbol{X'}_{it}$ is the vector (s) of control variables including year fixed effect and asset value, and $e_{it}$ is the error.

The regression results are shown in Table 12.

**Table 12 – Market power and risk exposure[126]**

***Loan interest rate elasticity and risk exposure***

| | Dependent variable: | | | |
|---|---|---|---|---|
| | *Loan loss ratio* | | *Subordinated debt asset to total asset* | |
| | (1) | (2) | (3) | (4) |
| | ***Pooled OLS with elasticity*** | ***Pooled OLS with HHI*** | ***Pooled OLS with elasticity*** | ***Pooled OLS with HHI*** |
| Own elesticity of loan interest rate | **0.16***<br>(0.02)** | | **0.00027***<br>(0.000071)** | |
| HHI | | **-0.0084***<br>(0.0033)** | | -0.00039***<br>(0.00012) |
| Asset value | 3.87e-10***<br>(6.44e-11) | 5.50e-10***<br>(1.83e-10) | 8.98e-12***<br>(1.65e-12) | 2.41e-11***<br>(6.21e-12) |
| Year2013 | 0.052*<br>(0.029) | 0.13***<br>(0.03) | -0.0006***<br>(0.0002) | -0.0005***<br>(0.0002) |
| Year2014 | 0.02<br>(0.03) | 0.05*<br>(0.03) | -0.0003*<br>(0.0002) | -0.0003*<br>(0.0002) |
| Year2015 | 0.016<br>(0.027) | 0.014<br>(0.029) | -0.0003<br>(0.00018) | -0.0003**<br>(0.0002) |
| constant | 0.45***<br>(0.09) | 1.31***<br>(0.02) | -0.0002<br>(0.0004) | 0.001***<br>(0.0001) |
| *Adjusted R-sq* | 0.127 | 0.0057 | 0.043 | 0.0577 |
| *N* | 6429 | 6429 | 6429 | 6429 |
| F (5,6429) | 37.35 | 8.132 | 17.73 | 13.52 |

[126] Table 12 shows the results of the estimation regarding the relationship between loan interest rate elasticity and risk exposure. More specifically, the statistics are the results of Ordinary Least Square estimation on specification (7). The numbers in the parentheses are the robust standard errors. Data sources: Bank Scope Database. Computed by the author.

As shown in the estimation result above, since I am using the absolute value of elasticities, if the own interest rate elasticity is high, then it means the market power of this bank is low and vice versa. I have done two separate regressions by using the HHI and the elasticity as market share measurement. The results of both are substantially aligning with each other: When the market power of a bank goes up, the risk exposure of that bank will go down. This conclusion is perfectly aligning with the conclusion in Salas and Saurina (2003) though the dataset that they use is the Spanish bank dataset. That is to say, banks with larger market power are more risk- averse. However, this doesn't explain the rationale: Does being risk-averse make them more discreet and therefore have better performance? Or it's vice-versa? More research must be conducted before these questions can be answered.

Based on the results of the tests here using both the elasticity and HHI as the measurement of market power, we fail to reject H4.

### *2.8.2 Market power and financial structure*

During the last two decades, many researchers have noticed the mutual relationship between the firm strategy for labor and the good market and their strategy of financing. Phillips (1995) finds that the capital structure of firms is highly correlated with industry output. They also find that the price in the final good market will be significantly changed once the major firms in those industries change their financial leverages. This paper essentially points out that the firm's financial strategy can indirectly make impacts on the good/services market and that there exists a strong connection between the goods/services market and the capital market. Chevalier (1995) uses the supermarket industry data to do empirical research and finds out that the announcement of a leveraged buy-out (LBO) increases the expected future profits of a firm's product-market rivals and that the presence of LBO firms encourages local entry and expansion by rivals. Bifid???, this paper is saying a firm’s strategy has certain relationships with industrial structures. Later on, MacKay and Phillips (2005) have shown that the firm’s financial strategies also highly rely on its position in the industry, and that is especially true for the competitive industries. They use the capital-labor ratio as a measure of the position of the firm in the market, and they find out that firms with capital-labor ratios close to the industry median (high natural hedge) useless financial leverage than firms that depart from the industry median capital-labor ratio (low natural hedge), as predicted by Maksimovic and Zechner (1991).

Based on the literature mentioned above, it’s clear that a firm’s financial strategy, including financial leverage, is highly dependent on the industry structure and its position in the industry. Therefore, I use the elasticity of the loan interest rate as the measurement of the firm’s market power and the position in the industry. By using the empirical framework developed by Phillips (2005), I test whether the financial leverage adopted by a firm in the banking industry also has a correlation with the firm’s position in the industry, which is measured by the market power. According to the traditional management theory and applied game theory literature (e.g., Aumann and Hart, 1992, Wagner, 1988), the firms which have largermarket power should have more bargaining power with their supplier and customer. Therefore, the firms that have lower elasticity of loan interest rates, which implied that largermarket power in the loan market, will likely to be able to borrow money at a lower interest rate from the money suppliers and I expect that the financial leverage is negatively correlated with the elasticity of loan interest rate.

An estimation of the following equation is conducted:

$$y_{it} = a + \beta * MP_{it} + \boldsymbol{X'}_{\boldsymbol{it}}\gamma + e_{it} \tag{8}$$

Where $y_{it}$ denotes the dependent variable which denotes the measurement of the financial structure of the bank, namely the financial leverage, $MP_{it}$ denotes the market power which is measured either by HHI or Own- Interest rate elasticity, $\boldsymbol{X'}_{\boldsymbol{it}}$ is the vector (s) of control variables including year fixed effect and asset value, and $e_{it}$ is the error.

The estimation follows a simple reduced-form regression and the framework is the same as the framework in 2.8.1. The estimation results and related statistics are shown in Table 13.

**Table 13 – Market power and financing structure**[127]

***Loan interest rate elasticity and financial leverage***

| | *Dependent variable:* | |
|---|---|---|
| | *Financial leverage* | |
| | (1) | (2) |
| | ***Pooled OLS with elasticity*** | ***Pooled OLS with HHI*** |
| Own elesticity of loan interest rate | **0.35** (**0.28**) | |
| HHI | | **0.08***** (**0.02**) |
| Asset value | -1.07e-09** (4.47e-10) | -4.15e-09 (1.16e-09) |
| Year2013 | 0.77** (0.32) | 0.94*** (0.25) |
| Year2014 | 0.55* (0.30) | 0.64** (0.27) |
| Year2015 | 0.17 (0.19) | 0.19 (0.19) |
| constant | 11.16*** (1.41) | 13.03 (0.14) |
| N | 6429 | 6429 |
| R-sq | 0.0104 | 0.0038 |
| F (5,6429) | 8.35 | 5.85 |

As we can see from the table above, for the leverage story, the HHI and the elasticity are telling different stories. In this analysis, financial leverage is defined by the equity value/ total asset. If we look at the estimation results by using HHI as the measurement of market power, it's saying that the banks with larger market power are having more equity

[127] Table 13 shows the results of the estimation regarding the relationship between loan interest rate elasticity and financial leverage. More specifically, the statistics are the results of Ordinary Least Square estimation on specification (8). The numbers in the parentheses are the robust standard errors. Data sources: Bank Scope Database. Computed by the author.

financing. However, in the regression of using the elasticity as the measurement of market power, the results are saying that there do not exist significant differences between the banks with bigger market power and the banks with smaller market power. Even if the estimator was significant, it's also telling us that the banks with more market power intend to finance more with debt. The estimation results given by the regression of using the elasticity as the measurement of market power is closer to our intuition.

With the limited information and data here, I can't figure out which version is better between the HHI story and the elasticity story. However, this might be another interesting topic for future research.

## 2.9 Management and policy implications

As discussed, the key findings of this paper provide management implications to the managers in banks and policy implications to the monetary policymakers, and those implications include the following:

1, Interest rates are still the most significant factor when customers are considering a loan. However, there are other factors that customers take into account. Customers also take the efficiency and helpfulness of employees into account. Therefore, banks can avoid price competition through improvements in service quality.

2, Larger banks are less sensitive to monetary policies that adjust the interest rates but small banks are much more sensitive. Because of that, managers in small banks should leverage the interest rate derivatives to hedge this kind of interest rate risk.

3, The Federal Reserve should consider the unexpected squeezing effects that the monetary policies may incur. The adjustments in interest rates might flush the small banks out and make the large banks more concentrated.

## 2.10 Concluding remarks

In this paper, I follow Berry et al. (1995) and Nevo (2000) to build up a structural model to estimate the loan demands of commercial banks in the U.S. I present the overall information and the importance of the banking industry and the American loan market at the beginning. Then, I build up the choice model based on the previous literature regarding structural models and bank customers' reference. I also propose the instrumental variables to solve the endogenous problem of loan interest rates. After getting the estimators, I compute the loan interest rate elasticities, based on which, I discuss the possible impacts from the monetary policy on the structure of the commercial banking industry by looking at the possible share changes of the largest bank and the smallest bank.

The key findings of this paper include the following: 1, the Interest rate of the loan is the key element in the demand system. Larger banks seem to have a smaller interest rate elasticity and smaller banks seem to have a larger interest rate elasticity. 2, Other than the interest rates, efficiency and helpfulness of staff are also factors that enter the customers' minds when choosing banks. 3, Monetary policies may have unexpected impacts on the structure of the commercial banking industry. 4, Banks with larger market power have less risk exposure. However, for the financial leverage hypothesis, the elasticity story is supporting the classic hypothesis that firms with higher market power are more likely to finance more by using liability while the HHI story is supporting the opposite.

There are several limitations to this paper. First, this paper strictly follows the BLP-type model and incorporates all the drawbacks of the BLP-type model. Secondly, there might exist a measurement error problem in the sense that the indicators and proxies are not the perfect indicators/proxies. For instance, using goodwill as the measurement for reputation might not be sufficient. However, due to the limited information that the banks disclose and the limited information that our dataset contains, goodwill might be the best proxy at this point in time. Thirdly, when doing GMM estimations, the objective function value is somewhat huge. This may indicate the model setup, and fitting is not good enough. This could be overcome by adding more variables or changing the optimization methods but might need much more computational power, which my devices don't have. Last but not least, the BLP estimation is using a small sample to avoid computational complications and downfalls. Though, as justified above, this shouldn't introduce any bias, this is still a limitation of this paper. However, this should not be seen as a flaw of this paper since it's a technical problem of the BLP model (the algorithm can be improved, and then the convergence property can be enhanced for large samples), and it's a hardware problem of the computer[128].

Due to the existence of the shortcomings mentioned, there are more that can be done in the future. Additionally, I noticed that the bank's stock price return is somewhat correlated with the own-interest rate elasticity. This may be an interesting inspiration for future asset pricing research.

[128] The processors are not fast enough and the RAM is not huge enough for simulation storage.

# CHAPTER 3. WHAT ARE THE DETERMINANTS THAT MAKE CHINESE FIRMS GO PUBLIC OVERSEAS: AN EMPIRICAL INVESTIGATION

## 3.1 Introduction

### *3.1.1 Institutional backgrounds of Chinese IPO*

The term Initial public offering (IPO) refers to a type of public offering in which shares of a company are sold to institutional investors that, in turn, sell to the general public, on a securities exchange, for the first time. The first modern IPO occurred in March 1602 when the Dutch East India Company offered shares of the company to the public in order to raise capital[129]. Initial Public Offering (IPO) has always been a focus in the finance field since then. Relative to liability financing, the expenses of equity financing are lower in most cases.[130] Equity financing has tons of benefits in addition to the relative inexpensiveness: It provides bargaining power for firms with banks[131]. It also increases the liquidity of the firms[132] and it enlarges the set of potential investors of the firms, which may enhance the corporate governance of the firms.[133] Therefore, at present, many Chinese firms go public and some of them go public overseas.

---

[129] Ammendola.(2000) "Devil Take the Hindmost: A History of Financial Speculation", Chapter 1:"This Bubble World": The Origins of Financial Speculation.

[130] Gaud et al. (2007) uses the European firms' data to show that the firms are not willing to borrow debts if they have other alternatives of financing. They imply that the cost of debt financing is higher than the cost of other financing alternatives.

[131] Rajan (1992) shows that IPO is more likely for companies paying higher rates and after IPO those firms experience decrease in borrowing rates

[132] Zingales (1998) shows that liquidity of a firm increases after IPO.

[133] Merton (1987) shows that after IPO diffuses stock ownership of firms.

Until the end of 2012, there are 2494 Chinese firms that are publicly traded domestically[134], and there are 261 firms that are publicly traded in the U.S at either the New York Stock Exchange (NYSE) or NASDAQ.[135] Those numbers show that the portion of Chinese public-traded firms willing to be cross-listed in the U.S is not small. In fact, the Chinese local equity markets unusually offer a huge premium when the firms IPO locally. However, there is still a relatively large portion of Chinese firms that IPO in the U.S. in spite of the high premium offered by Chinese domestic investors.[136] This phenomenon has been documented by many scholars (e.g. Chan et.al, 2008). Up to this point, there exist very few theoretical or empirical research providing satisfactory answers to this puzzle.

### *3.1.2 Key findings and contributions*

In this paper, I explore the external and internal determinants that make Chinese firms go public overseas or domestically in spite of the premium offered by the local capital market. The key methodology of this paper is to use cross-sectional discrete choice models to test the hypotheses regarding this issue. Also, the ex-post analysis will provide robustness checks for the ex-ante analysis results. This paper finds that huge structural changes have been observed in Chinese firms' preferences over IPO locations. Generally, small firms with higher growth and facing the borrowing constraints are more willing to IPO overseas. The firm information disclosure and firm transparency level after IPO are

[134] Based on the data provided by Sina Finance:
http://finance.sina.com.cn/worldmac/indicator_CM.MKT.LDOM.NO.shtml
[135] Based on the data provided by NYSE and NASDAQ official website. The number refers to the total number of firms that are participating in all level ADR programs.
[136] The premium here refers to the higher willingness to pay for a stock by the investors due to the positive perspectives.

key and robust determinants when firms making IPO choices: the firms that have inadequate information disclosure and shareholder protection are more likely to IPO overseas.

This paper, in spite of the limitation of data, provides a good framework for the analysis of the cross-listing behaviors of Chinese firms. Besides, the conclusions obtained from the analysis in this paper answers the question of why Chinese firms are willing to give up the premium offered by local investors for the sake of IPO overseas to a great extent.

## 3.2 Literature review

### *3.2.1 Literature on cross-listing*

Under the background of globalization, cross-listed stocks have become much more prevailing. Cross-listing issue has been a focus in the international finance area because cross-listing stocks provide many cases in which the law of one price doesn't hold (e.g Eun et al., 1988), especially in the situation where the local market trading time overlaps with the cross-listing markets. Thus, many scholars focus on pricing and price movements. For instance, Werner et al. (1996) analyze the volume, spread, and volatility of cross-listed stocks and find that during the overlap trading time period, the cross-listed stocks have significantly higher trading volatility and law of one price doesn't hold. Following that methodology, Pascual et al. (2006) analyze the price discovery process of securities that trade at multiple markets with trading sessions that totally or partially overlap. They find that the NYSE in the price discovery process of the Spanish cross-listed and find that the SSE (*Spanish Stock Exchange)* trade-related shocks account for a large portion of the long-run variance of the Spanish cross-listed stocks, which can influence the risk premiums for these stocks. Chan and et al. (2008) investigates the relation between cross-listing in the United States and the information environment of non-U.S. firms. They find that firms that cross-list their stocks on the U.S. exchanges have greater analyst coverage and increased forecast accuracy than firms that are not cross-listed, and that firms that have more analyst coverage and higher forecast accuracies have higher market values. Aside from the empirical work, theoretical work has been done regarding the asset pricing issue of cross-listed stocks. Eun et al. (1987) derive an asset pricing model showing that the value of the stock is not only correlated with the local market but also correlated with the foreign markets where the stock is cross-listed . By following this theoretical framework, Eun et al. (1988) perform an empirical test in which they find that the international listing of security should, in general, accompany a reduction in its expected return.

### *3.2.2 Literature on IPO and cross-listing behaviors*

A set of works have been done in a more focused way that they concentrate on why firms would like to IPO or be cross-listed overseas. Chan and et al. (2008)'s finds that cross-listing can increase the analysts' coverage and hence, increase the value of the firm. Thus, the increase in analysts' coverage and increase in value will be two possible reasons for cross-listing overseas. Baker et al. (2002) also investigate what cross-listing brings to firms. They find that cross-listing firms in NYSE and LSE experience decreases in the cost of equity capital. Amir (2003) proposes that from the perspective of capital market regulation, cross-listing should incentivize more self-disclosures by firms and can improve corporate governance. Some firms actually become cross-listed for the purpose of shareholder protection. Another empirical work that comprehensively examines the reason for cross-listing by Sarkissian et al. (2004). They find that the cultural effects, disclosure level of information and tax, the degrees of correlation and integration between the home market and the cross-listing foreign market, and legal environments are the key determinants making firms choose to be cross-listed in other countries. Doidge (2009) has discovered that the private benefits of control do have an impact on the decision of cross-listing or IPO overseas from a purely empirical perspective.

Unfortunately, the literature on why companies IPO in foreign markets is quite limited. Blass et al. (2001) adopt the methodology of combining ex-ante and ex-post analysis and find out that on firm-level, R&D intensity, the firm size, revenue, and industry categories have significant impact on the probability of local companies IPO in the U.S. On the other hand, also very few research put their focus on the Chinese firms' behaviors regarding going public in other countries. Sami et al. (2008) leverage the data of Chinese firms and find out that information asymmetry risk will decrease, and firm disclosure level will increase after IPO overseas or cross-listing.

### *3.2.3 Literature regarding impacts from the government on IPO behaviours*

Another steam of works that I have looked at includes the literature about the impacts of government and market information asymmetry on international trade and finance. Doidge et al. (2004) find that eliminating the information asymmetry makes the value of the stocks increase. Also, the government that has a higher regulatory standards for its financial market can help add values to the stocks listed in its financial market. Levchenko (2007) suggests that in international trade, the quality of the government has an impact on the trade flows in the good market. Similar research has been done by Berkowitz et al (2006) which find that the trade flow in good markets between countries is influenced by the quality of law and regulations made by the governments on both sides.

## 3.3 Conceptual framework

### *3.3.1 Theoretical framework*

As pointed out in the literature review section, most of the previous research related to oversea IPO is empirical. Until now, we don't have a complete and well-developed theoretical framework to explain the reasons why many firms choose to make IPOs in other countries or regions, giving up the premium offered by the local investor. In order to provide solid theoretical framework for the empirical analysis, I follow a classic theoretical framework from an industrial organization developed by Bresnahan and Reiss (1991) which develops a theoretical framework for binary choice of entry and exit by firms that is quite similar to the binary choice of IPO locally and IPO overseas in the sense that the firms are choosing the optimal strategy to maximize the prospective profit. In addition to Bresnahan and Reiss (1991), I also use the framework developed by Aslan and Kumar (2017), which explains the rationales behind individual decisions of entrepreneurship under the context of increasing import exposure.

In the binary choice setting, the firms are trying to compare the prospective utility from each alternative, which can be measured by economic benefits that they can obtain from each alternative. Let the prospective utility be a function of two parts of economic values – the economic value that it can obtain by going public and the economic value added by the destination market.

Blass et al. (2001) talk about the benefits of IPO for firms in his ex-post analysis which identifies the reason for firms to go public. Essentially, the firms have quite similar benefits when going public. They all experience an increase in ROA, increased revenue, and higher growth rate. However, those effects are not homogeneous, and it varies across industries and firms. As for the benefits of IPO overseas, as pointed out by the previous literature review , the main determinants of the benefits for IPO overseas include the benefits and costs of information disclosure, the market premium, institutional and government protection and the firm-specific effects which are determined by firm characteristics. Therefore, we can model the choice behavior as the following:

$$U_{ij} = u_{IPO}(X_i) + u_j(InfoD_i, IGP_i, Premium_i, X_i) \quad (1)$$

where $U_{ij}$ stands for the total indirect utility that can be obtained from IPO choice j by firm i. $u_{IPO}(X_i)$ denotes a non-stochastic function of benefits of going public given the firm characteristics $X_i$ of firm i. $u_j(InfoD_i, IGP_i, Premium_i, X_i)$ is the non-stochastic function of benefits of going public in the j market by firm i given its information disclosure level $InfoD_i$, Institutional and government protection that they are seeking $IGP_i$ and the market premium it's going to get $Premium_i$ and firm-specific effects which are determined by the firm characteristics $X_i$.

The prospective indirect utility of IPO locally is

$$U_{i0} = u_{IPO}(X_i) + u_0(InfoD_i, IGP_i, Premium_i, X_i) \quad (2)$$

and the prospective indirect utility of IPO overseas is

$$U_{i1} = u_{IPO}(X_i) + u_1(InfoD_i, IGP_i, Premium_i, X_i) \quad (3)$$

From the perspective of the firms, when making decisions over these two choices, they are comparing $U_{i0}\ with\ U_{i1}$. For specific firm i, since $X_i$ is the same for both indirect utility function, the utility from going public is the same for both choices. Then, comparing $U_{i0}\ with\ U_{i1}$ would simply become a comparison between $u_0(InfoD_i, IGP_i, Premium_i, X_i)$ and $u_1(InfoD_i, IGP_i, Premium_i, X_i)$.

For the firms that have the IPO overseas, $u_1(InfoD_i, IGP_i, Premium_i, X_i) - u_0(InfoD_i, IGP_i, Premium_i, X_i) > 0$, which implies $u_1(InfoD_i, IGP_i, Premium_i, X_i) > u_0(InfoD_i, IGP_i, Premium_i, X_i)$. Therefore, the set of the firms that choose to IPO overseas will be the set of

$\bar{I} = \{i|U_{i1} > U_{i0}\}$, which is also equivalent to $\bar{I} = \{i|u_1(InfoD_i, IGP_i, Premium_i, X_i) > u_0(InfoD_i, IGP_i, Premium_i, X_i)\}$ (4)

As suggested by Fernald and John (2002) and Ammendola (2000), due to the lack of alternative of investment in China, the premium offered by the local market is huge if a firm IPO locally. On the contrary, since other markets don't have this problem and due to the equity home bias (Bernile et al, 2015), if a Chinese firm IPO in the markets other than the Chinese market, they are not likely to receive any premium. Assuming the premium money and the utility of the firm has a 1-to-1 relationship, and that the premium value has a linear relationship with other parts of the indirect utility, the set of the firms that choose to IPO overseas will become

$$\bar{I} = \{i|u_1(InfoD_i, IGP_i, X_i) > u_0(InfoD_i, IGP_i, X_i) + Premium_i\}. \qquad (5)$$

By looking at this set, for the firms which IPO overseas, the difference in the net information gain/loss plus the difference in the institutional and government protection plus the difference between the firm-specific gains between the two choices must overweight the premium offered by the local market.

In the equilibrium, the condition for firms to IPO overseas is

$$u_1(InfoD_i, IGP_i, X_i) \geq u_0(InfoD_i, IGP_i, X_i) + Premium_i \tag{6}$$

And can be transformed into an equivalent condition

$$u_1(InfoD_i, IGP_i, X_i) - u_0(InfoD_i, IGP_i, X_i) \geq Premium_i \tag{7}$$

In other words, firms and their controls (usually the boards) are trying to maximize their utility between the two IPO options by looking at the difference in informational cost and gains, the institutional and governmental protection that they are seeking and the firm-specific gain or loss between these two options. They are always trying to see whether those differences can overweight the premium: If yes, they will choose to IPO oversea, if no they will choose IPO locally.

Therefore, in order to compare $u_1(InfoD_i, IGP_i, X_i)\ and\ u_0(InfoD_i, IGP_i, X_i) + Premium_i$ estimating the random utility model can help to infer whether those determinants outweigh the premium.

### *3.3.2 Hypotheses to test*

Based on the theoretical framework in the previous section, the key hypotheses to test include the following:

**H1**: The information cost/gain, measured by tax-to-revenue ratio, is a factor that dominates the premium offered by the local market and makes firms IPO overseas

**H2**: The institutional and governmental protection in the oversea markets is one of the key factors that dominate the premium offered by the local market. With good governmental protection, firms are more willing to IPO in that market.

### *3.3.3 Empirical methodology*

#### 3.3.3.1 Random utility framework

As mentioned in the literature review, my methodology of analysis mostly follows McFadden's (1973) framework for discrete choice estimation, assuming the IPO decisions are made by the manager, board (if there is one) and shareholders of a firm. For simplicity, we assume they have the same preference and the same indirect utilities from different IPO alternatives. When they are deciding if they are going to IPO overseas or not, they are making the decision as a whole and have the utility function as

$$U_{ij} = X_i\beta + \epsilon_{ij} \tag{8}$$

Where $X_i\beta$ is the utility from generated based on the observed characteristics of firm i and $\epsilon_{ij}$ is the unobserved utility generated by another alternative or case-specific characteristics when choosing alternative j. When the firms are going to determine if they want to IPO overseas or not, the alternative set of j can be denoted as S={Overseas, Locally}. When they are determining in which market to do the IPO, their alternatives set becomes S'={U.S., Hongkong, Germany, U.K., Singapore, China mainland}

Thus, a rational decision-maker will choose between j by comparing $U_{ij}$. Assuming the probability of equal utilities equal to 0, the probability of choosing alternative j should equal to $\Pr(J = j) = \Pr(U_{ij} > U_{ik})$ where $k \neq j$. Then the probability of choosing alternative J depends on the difference between $U_{ij}$ *and* $U_{ik}$. By following the classical literature of discrete choice model (e.g. McFadden(1973)), when firms are making decisions about if they want to IPO overseas, the $\epsilon_{ij}$ is following a type-1 extreme value with distribution function $F(\epsilon) = \exp(-\exp(-\epsilon))$, the difference between $U_{ij}$ *and* $U_{ik}$ will follow a standard logistic distribution. When difference between $U_{ij}$ *and* $U_{ik}$ follows a normal distribution, then we will have a probit choice model. Overall speaking, the logit models have a better fit for the data[137].

### 3.3.3.2 Possible determinants and estimation framework

Though a few firms in our sample become cross-listed in China mainland or Hongkong, the information effects of cross-listing suggested by Werner et al (1996) and others should not be considered as the determinants that determine the IPO location of the firm. The reasons are the following: 1, the purpose of this paper is to see what the determinants are that drive the Chinese firms' IPO and raise capital in other countries. The focus of this paper is to explore the determinants that influence the IPO decisions of the firms instead of cross-listing decisions; 2, the effects of information capitalization and analyst coverage are proved to be more related to cross-listing versus IPO since newly IPO stocks usually can't be traded in two synchronized markets. At least there are no such

[137] Details and tests will be shown in the robustness check part.

stocks in my sample. Therefore, the analyst coverage and information coverage will not be assumed to have an impact on the IPO decisions of the Chinese firms.

Based on the theoretical framework and empirical results of Blass et al (2001), Sami et al (2008) and Sarkissian et al (2004), on the firm level, there are several key factors that may have impact on the firms' decisions to IPO overseas including R&D intensity, firm size, revenue, industry categories, firm information disclosure level, culture, location are the firm-level determinants for firms to cross-listing and IPO in other countries.

Aside from those factors proposed by previous scholars, three additional factors are supposed to have an impact on the choice of IPO. One is financial structure, the second variable is growth and the other one is capital expenditure. The reason why these factors should be included is that they are reflecting the borrowing constraints of the firms, as indicated by Pagano (1998). When different firms are confronting with different borrowing constraints, they may make different decisions regarding where to IPO since for some firms. IPO overseas may possibly help overcome the borrowing constraints they have. Also, the growth rate is a factor showing the status of a firm. Intuitively, the status of development will definitely influence the firm's IPO decision and choice since its strategy is based on the current status of development.

Based on previous researches mentioned in the literature review section regarding information asymmetry and government quality on international trade and finance, the government quality and information asymmetry in the market will create different effects on different firms and thus, different firms will make different IPO decisions because of that. I add in two alternative-firm determinants that may have an impact on the IPO decisions.

The first alternative-firm determinant is the firm transparency in the destination market. The information undisclosed can generate asymmetry by firms. However, on the other hand, if the information dispersing channel in that market is bad, then no matter how much information a firm discloses, there still might be huge information asymmetry for investors and the firms still be non-transparent. Therefore, the firm transparency in the destination market in the destination market should not be unified for all the firms. The firm transparency in the destination market is determined by the information disclosure by firms and the efficiency of the information dispersing channels simultaneously. Therefore, I interact these two variables obtain a new individual-specific variable.

The second alternative-firm determinant is the government's quality effect on firms. When firms IPO in other countries, basically they are “exporting” stocks, a financial good, to other countries. Legally speaking, the procedure of selling stocks and the terms of the shares will be subject to the law, regulation and government ruling of the destination market. However, this effect shouldn’t be uniform either since the size of the stocks that firms sell to foreign markets is different. In that sense, this effect is depending on the size of stocks issued to foreign investors when IPO, and I interact these two variables to make a new variable which is individual-specific.

#### 3.3.3.3 The empirical model and specification

To sum up, possible determinants for firms to IPO abroad include R&D intensity, firm size, revenue, industry categories, firm information disclosure level, culture, location, regulations of financial markets, correlations and integration degrees and exchange rates. Theoretically, it can be implied by the theoretical model:

$$\Pr(\mathrm{y}_i = \mathrm{J}|\ \mathrm{R\&D\ intensity} \ldots) = \mathrm{f}(\mathrm{R\&D\ intensity}_i + \mathrm{firm\ size}_i + \mathrm{industry\ categories}_i + \mathrm{firm\ information\ disclosure\ level}_i + \mathrm{culture}_i + \mathrm{location}_i + \mathrm{financial\ leverage}_i + \mathrm{growth\ rate}_i + capital\ expenditure_i + Information\ Dispersion_j * Information\ Disclosure_i + Size\ of\ Issuance_i * Government\ Quality_j) \qquad (9)$$

3.3.3.4 Proxies for independent variables

Since all the databases that are available to me do not contain any information regarding R&D intensity for locally IPO firms,[138] it won't be able to let me test whether R&D intensity has an impact on the decision of IPO. As for firm size, I use the total asset value as the proxy for firm size. This proxy is used by Blass et al (2001) and has been proved to be robust by Dang et al (2015). I will use the revenue of the previous year of the firm I because usually, an IPO is planned 2 years before the final IPO. Firms are not looking at their current revenue to make decisions. Thus, the revenue of the previous year instead of the revenue of the year of IPO is the consideration taken into account when making IPO decisions. For industry, I use a dummy to show if it's in IT industry since Blass et al (2001) shows that firms in the IT industry are more propended to IPO in the U.S.

It's extremely hard to measure the firm's information disclosure level, especially when I don't have complete corporate governance data before they IPO. A good and simple proxy is a firm's income tax according to Pagano (1998). Pagano (1998) uses it as a

[138] The data set doesn't have R&D expenses or patten information for most firms in the sample.

measure of loss of confidentiality and provides a solid argument for that. Thus, I believe this variable should be a good proxy for information disclosure. However, the absolute values of income taxes are not comparable between firms. In order to solve this problem, I compute the tax/revenue ratio to proxy for firm information disclosure level . Additionally, I think this is not enough for controlling the information disclosure since the low tax/revenue rate might also indicate high costs, thus I also control for the margin to control for the cost difference. The margin is computed by using net income before tax divided by total revenue. As for culture and location, I will use a dummy variable to distinguish between the firms that are on the coastal area or in the inner land area of China. The province defined as the coastal area is shown by the map in Appendix A. The reason for such division is that firms that are on the coastal area usually have a long history because they started earlier due to the policy of reformation and opening. They are more affected by foreign cultures and technologies relative to the inner land areas. So this dummy should be able to capture the culture and location effects. Financial leverage will be computed through total equity divided by total liability, which is the definition of a financial leverage ratio. Based on Pagano (1998), the growth rate will be indicated by the percentage change of revenue from the previous year to the IPO year.

As for the alternative– specific determinants, I use the internet user rate in each country of the year right before IPO to proxy for the information dispersion efficiency. I also calculate the income tax/ revenue ratio[139] and use the interaction term between these two to proxy for the firm transparency. I use the equal-weighted average of the index of

[139] I compute this ratio in order to avoid the multicollinearity issue. If I interact the margin and tax/revenue ratio separately with the internet user ratio, there will be a multicollinearity problem in the logit model.

the governance indicators of the destination country to proxy for the government quality. Since I don't have the data for the size of IPO issuance, I use the equity size as the issuance size. This should be a reasonable proxy because usually, the IPO issuance size is proportional to the equity size.

The reason why I use the data of the year right before IPO is that these kinds of data are the only pre-IPO data I can obtain for Chinese firms. Many Chinese firms don't publish quarterly data before they go public. Firms are usually planning IPO two years ahead of their actual IPO data on average[140]. Once they plan IPO, they will try to adjust their financial indicators in order to fit their IPO plan. Thus, the data of the year right before the IPO year should be reflecting the latest information before IPO and this information should be enough to reveal their primary choice of alternatives.

Based on the requirement of GAAP(Generally Accepted Accounting Principles) and IFRS (*International Financial Reporting Standards)*[141], I convert the assets, capital expenditures into Chinese currency based on the exchange rate of the last trading day of the year before IPO.

Then, what I am going to estimate will be

$$\Pr(y_i = \mathrm{J} |\ \mathrm{ASSET} \ldots) = \mathrm{f}(\ \beta_1 \mathrm{ASSET}_i + \beta_2 \mathrm{D1}_i + \beta_3 \mathrm{Tax/Rev}_i + \beta_4 \mathrm{D2}_i + \beta_5 \mathrm{LEVERAGE}_i + \beta_6 \mathrm{GROWTH}_i + \beta_7 \mathrm{CAPEX}_i + \beta_8 Margin_i + \beta_9 ICTAX_i * INET_j + \beta_{10} Equity_i * Gov_j) \tag{10}$$

[140] Chi, Jing, and Carol Padgett(2005) points out that the average cycle for Chinese firms to IPO is in the range of 1-3 years and the average is about 2 years.
[141] Though there are some difference between GAPP and IFRS, they are not difference when it comes to balance sheet item conversion and cash flow conversion.

$\mathrm{ASSET}_i$: The total asset of the year before IPO of firm i

$\mathrm{D1}_i$: The dummy variable showing if the firm is in the IT industry or not. If yes, it equals to 1, otherwise, it equals to 0.

$\mathrm{Tax/Rev}_i$: Income tax/ revenue ratio of the year before the IPO of firm i.

$\mathrm{D2}_i$ : The dummy variable showing if the firm is in a province defined as the coastal area. If yes, it equals to 1, otherwise, it equals to 0.

$\mathrm{LEVERAGE}_i$: The financial leverage ratio of the year before IPO firm i.

$\mathrm{GROWTH}_i$: The revenue percentages change from revenue of the previous year to the IPO year of firm i.

$\mathrm{CAPEX}_i$: The capital expenditure of the year before IPO of firm i

$Margin_i$:.The net income of the year before IPO before tax divided by total revenue of firm i.

$ICTAX_i$: Income tax/ net income of the year before the IPO of firm i.

$INET_j$: Internet user rate of the year before IPO in country j

$EQUITY_i$: The equity value of the year before IPO of the firm i,

$GOV_j$: Internet user rate of the year before IPO in country j

### *3.3.4 Estimation strategies*

#### 3.3.4.1 Binary response model

According to the framework built up to explore the determinants that make Chinese firms go public overseas, binary logit and probit regression will be ideal for estimations. Both results of binary logit regression and probit regression will be reported so that they can serve as robustness checks for each other. Probit and logit regression are well defined and frequently used models for binary response data and have been used a lot by famous researchers like Pagano (1998).

#### 3.3.4.2 Multinomial model

Based on the framework of the analysis proposed above, multinomial logistic (MNL) regression will be a good fit for my analysis. However, when we use the multinomial logit regression is used to model choices, which is exactly what I am going to do, it relies on the assumption of independence of irrelevant alternatives (IIA). Thus, it will be necessary to test if the IIA property by using the method proposed by Hausman and McFadden (1984). The reason why I think testing the IIA property for these two alternatives is necessary is that the market correlation between China mainland and Hongkong is very high, which indicates that these two markets may be highly integrated. We are not testing all the alternatives on the IIA property because the rest alternatives are distinct in terms of geographic location, local political environment, and information transparency, as indicated by the data. Therefore, the rest alternatives shouldn't become substitution of each other and hence, shouldn't violate the IIA property in general.

In case the IIA property doesn't hold, the nested logit model will be used in place of the multinomial logistic model. The nested logit model can relax the IIA assumption and hence, can help reduce the bias if IIA property doesn't hold.

## 3.4 Data and variables

### *3.4.1 Samples and time periods*

The data sources of the dataset used in this paper include CSMAR database and RESSET database, which is the most comprehensive database of Chinese public-traded firms. Our case-specific variables data of Chinese firms, including the asset value, financial leverage, income tax of the previous year, the revenue of the previous years, capital expenses of the previous year, industry category and headquarter location are mainly obtained from CSMAR and RESET database.

Compustat database is also leveraged to calibrate the data of foreign IPO firms. Also, it's used as a complementary database for the firms that have missing information from RESET database.

As for the alternative specific variables, the internet user data comes from the World Bank database used by Kaufmann et al. (2010). Government quality data is obtained from the Worldwide Governance Indicators database. It mainly contains 6 indicators, which are Voice and Accountability indicator, Political Stability and Absence of Violence/Terrorism indicator, Government Effectiveness indicator, Regulatory Quality indicator, and Rule of Law Control of Corruption indicator.

Before getting to the reasons for the time period selection, I need to make the goal of my research clear. In this paper, I am particularly concerned with the reasons that make the Chinese firms IPO overseas. Many firms choose to IPO at one location and then choose to cross-list their stocks at another location. For those firms, I only looked at their initial offering behaviors and I don't really take their cross-listing behaviors into account. For example, a firm originally IPO in the U.S and then cross-list their stocks in Hongkong, I only look at their data when they are preparing for IPO in the U.S. The CSMAR database and RESSET database have complete information about time when the stocks of a firm firstly become available, which allow me to know the IPO date of the firm and classify it into the corresponding year category.

Secondly, I am particularly interested in the IPO behaviors of the firms who choose to IPO in the U.S. Thus, I particularly care about the sample size of firms choosing this alternative. The behaviors of these firms are particularly interesting since U.S. has a more rigorous information disclosure requirement than China, which means it may incur more information costs if they choose to IPO in the U.S. Given the fact firms IPO in other countries have already given up the high premium offered by the local investors, it will be interesting to see what the internal factors are that drive them to incur more costs.

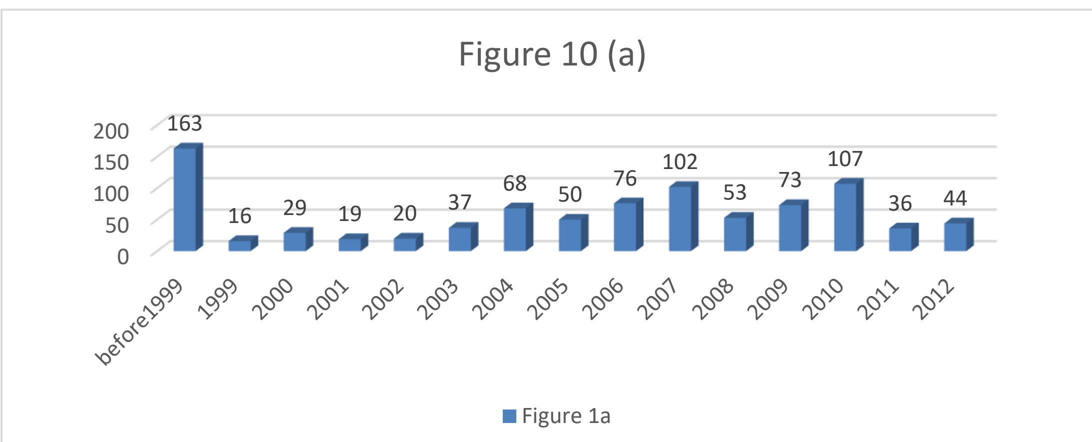


**Figure 10(a) – It shows the number of Chinese firms that become cross-listed or IPO in other countries or areas in a specific year or period.**

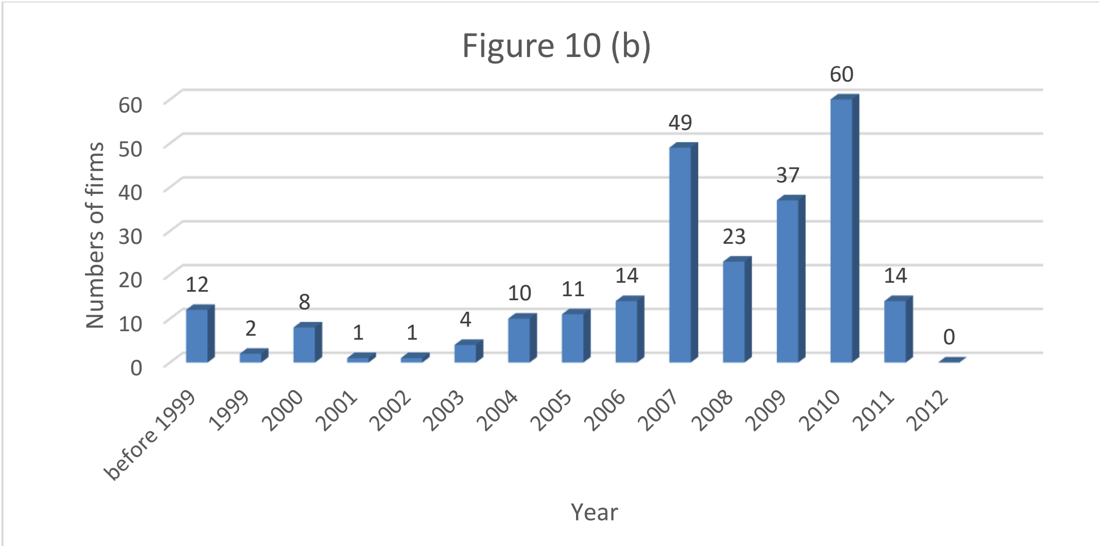


**Figure 10(b) – It shows the numbers of Chinese firms that become cross-listed or IPO in the U.S. in a specific year or period.**

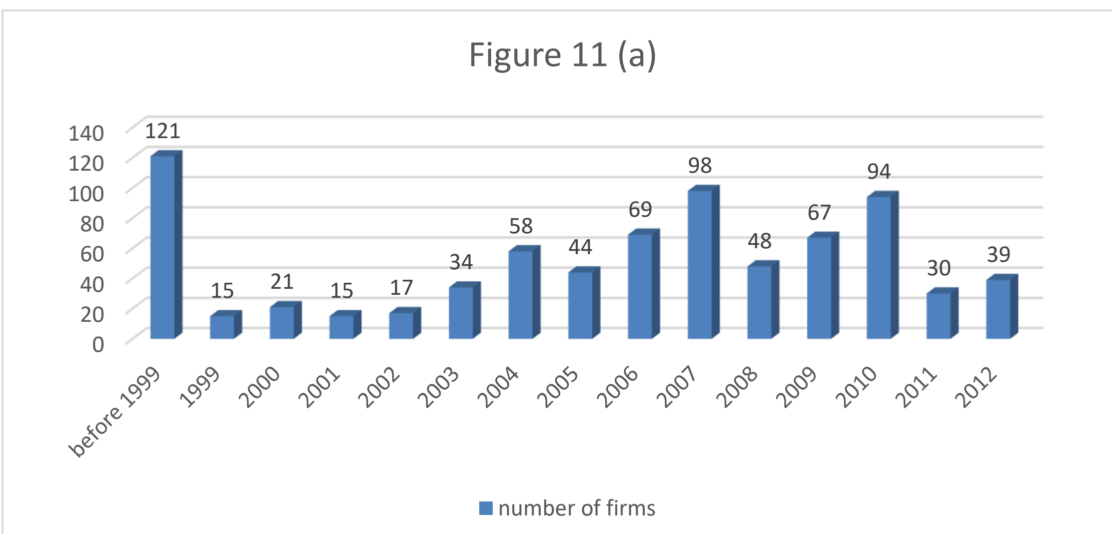


**Figure 11(a) – It shows the numbers of Chinese firms that IPO in other countries or areas in a specific year or period.**

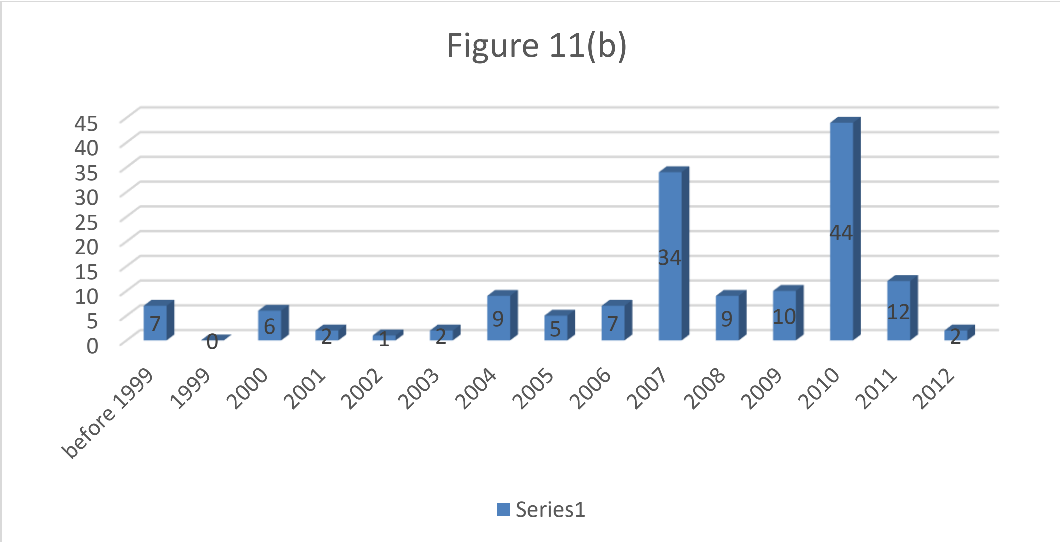


**Figure 11(b) – It shows the numbers of Chinese firms that IPO in the U.S. in a specific year or period.**

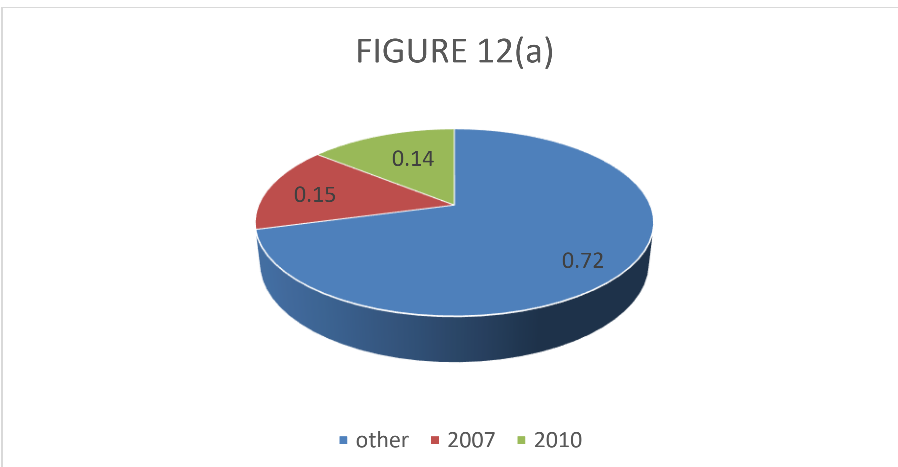


**Figure 12(a) – It shows the shares of Chinese firms that IPO in other countries or areas of a specific year or period.**

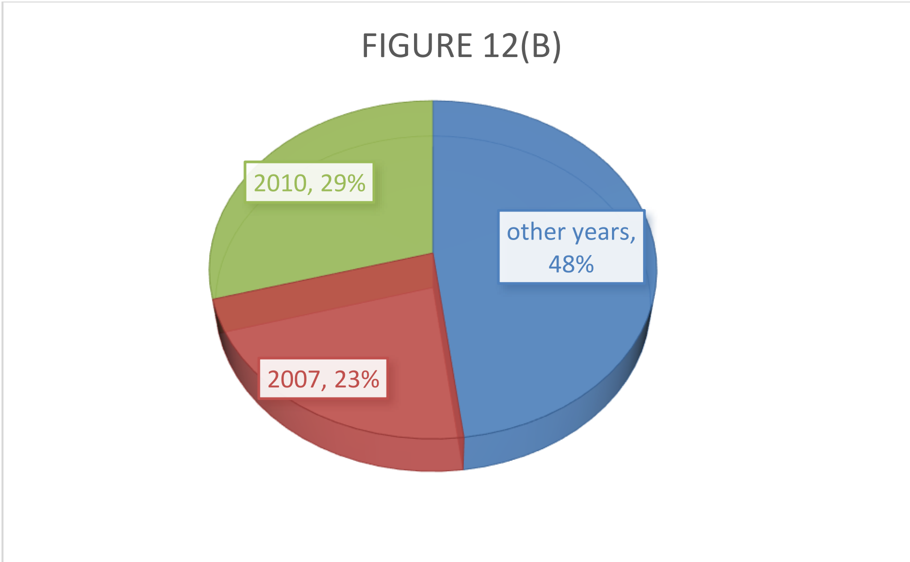


**Figure 12(b) – It shows the shares of Chinese firms that IPO in the U.S. of a specific year or period.**

Based on the summary of CSMAR database, most of the Chinese firms decide to be cross-listed or IPO in other countries or areas in 2007 and 2010, as shown in Figure 10(a). Also, most of the firms that IPO in the U.S. in the year 2007 and the year 2010 as suggested in Figure 11(b). The database doesn't contain the details showing which American depositary receipt (ADR) level the firms are, by looking at the NYSE website and, we still can identify the cross-listing firms. However, since in this paper, our concern is not the cross-listing location choices made by the Chinese firms. Instead, our focus is on the IPO choices and IPO behaviors made by Chinese firms. Then, I need to figure out in what years there are most firms IPO in other countries. Suggested by Figure 11(a) and Figure 12(A), the year 2007 and the year 2010 have the most firms that IPO overseas. Also, as shown in Figure 11(b) and Figure 12(b), most of the ADR level 3 participants IPO in the U.S during 2010 and 2007. Around 30% of the ADR level 3 program participants IPO in 2010. In the meanwhile, there are also a number of firms IPO domestically.

Therefore based on the purpose of this paper, the best sample period will be the data of the year 2010 and the year 2007. Another reason for choosing 2010 and 2007 as the sample period is that between 2007 and 2010, there was a severe global financial crisis. By making comparisons between those two years, we may identify the changes in the IPO behaviors before the crisis and after the crisis, which may bring up some policy implications.

The reasons why I am not using the panel data that covers the whole period include the following: Firstly, many of the pre-IPO data of the firms that IPO in the U.S before 2004 are largely missing. Secondly, some firms become delisted soon after the IPO. Later on, some of those firms which become delisted become listed in another market or the same market again. In that sense, if I use panel data to do the analysis, we have the duplication issue, which might create biases.

### *3.4.2 Data screening*

First, I delete the firms that do not have complete information regarding the key variables that we are concerned about. If any case-specific variable or the alternative-specific variable that we are concerned is missing from CSMAR database, RESSET database, and Compustat database for the years 2010 and 2007, this firm will be deleted from the sample since they are censored data. Deleting these firms will not generate any selection bias as long as the process generating missing data is random. Also, according to

McFadden (1981), endogenous sampling under the multinomial logit framework is not an issue if we are not concerned about the constant terms.

Second, I delete the outliers. The criteria for being an outlier is to have an asset value larger than 1 trillion CNY. The reason why I remove these firms from the sample is that some of the firms with such huge assets are government-owned enterprises. The decision of IPO for these firms will involve extremely complicated political considerations, which are very difficult to control. Other firms with a total asset value exceeding 1 trillion CNY are obviously outliers or data mistakes since the total asset value of all publicly traded firms in the real estate industry is 347 billion CNY by the end of 2013. Still, I assuming the process creating outliers/ data mistakes is exogenous, and then no bias will be introduced into the estimation due to the elimination of outliers.

### *3.4.3 Final sample*

A statistical summary is given below regarding the final sample in Table 14. The final sample of the year 2010 contains 350 firms that IPO in China, 39 firms that IPO in the U.S, 10 firms that IPO in Hongkong and 6 firms that IPO in Singapore and only one firm that IPO in Germany. Unfortunately, the data level of the firms that IPO in the U.K. is not complete and many key variables are missing. Thus, it’s not possible to include the one and only firm that IPO in the U.K. in 2010. I believe this will not cause serious bias since there was only one Chinese firm that IPO in the U.K. in 2010. Compared to the sample that we have, one firm is sort of trivial, which shouldn’t affect the results.

**Table 14(a) – Summary statistics of the year 2007**

**Year 2007**

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| **IPO overseas** | | | | | |
| leverage | 44 | 2.247136 | 3.938785 | 0.046192 | 21.22859 |
| asset | 44 | 21617.81 | 107475.8 | 56.73 | 704513 |
| equity | 44 | 19728.44 | 102652.7 | 30.37 | 672819 |
| growth rate | 44 | 0.820571 | 1.850948 | -0.56838 | 11.94016 |
| margin | 44 | 0.248499 | 0.188957 | 0.022394 | 1.001837 |
| tax/revenue | 44 | 0.039953 | 0.04782 | -0.0117 | 0.202738 |
| capital expenditure | 44 | 279.9556 | 822.3316 | 0.25 | 5206 |
| retain earnings | 44 | 537.7094 | 1284.097 | -665.923 | 7001 |
| culture | 44 | 0.863636 | 0.347142 | 0 | 1 |
| industry | 44 | 0.136364 | 0.347142 | 0 | 1 |
| **IPO locally** | | | | | |
| leverage | 120 | 1.237636 | 1.352518 | 0.04717 | 9.80275 |
| asset | 120 | 19239.27 | 102297.5 | 15.0712 | 815144 |
| equity | 120 | 7138.626 | 53236.22 | 7.550529 | 567595 |
| growthrate | 120 | 0.408653 | 0.562251 | -0.30901 | 4.081981 |
| margin | 120 | 0.141111 | 0.094122 | 0.014554 | 0.441227 |
| taxpropo | 120 | 0.025011 | 0.025033 | -0.01053 | 0.124848 |
| capex | 120 | 450.2857 | 2397.209 | 0.049548 | 23935 |
| retain2 | 120 | 715.2697 | 4319.881 | -84.8247 | 43092 |
| culture | 120 | 0.775 | 0.419333 | 0 | 1 |
| industry | 120 | 0.258333 | 0.439554 | 0 | 1 |

**Table 14(b) – Summary statistics of the year 2010**

| **Year 2010** | | | | | |
|---|---|---|---|---|---|
| Variable | obs | Mean | Std. Dev. | Min | Max |
| **IPO overseas** | | | | | |
| leverage | 56 | 2.702842 | 2.403061 | 0.085675 | 9.334857 |
| asset | 56 | 3336.223 | 5859.878 | 225.31 | 40933.3 |
| equity | 56 | 1601.515 | 1857.656 | 37.40055 | 10530.39 |
| growth rate | 56 | 2.020054 | 9.546223 | -0.49197 | 71.69666 |
| margin | 56 | 0.055956 | 0.551496 | -3.70081 | 0.504815 |
| tax/revenue | 56 | 0.01764 | 0.083334 | -0.55524 | 0.124704 |
| capital expenditure | 56 | 319.7188 | 1172.53 | 0.17 | 8140.76 |
| retain earnings | 56 | 2098.064 | 11309.83 | -490.135 | 81348.6 |
| culture | 56 | 0.839286 | 0.370591 | 0 | 1 |
| industry | 56 | 0.160714 | 0.370591 | 0 | 1 |
| **IPO locally** | | | | | |
| leverage | 353 | 1.816017 | 2.476086 | 0.040157 | 36.33209 |
| asset | 353 | 30117.87 | 476817.5 | 13.42975 | 8882588 |
| equity | 353 | 1694.266 | 18445.32 | 7.091014 | 342925 |
| growth rate | 353 | 0.264375 | 0.571286 | -0.54646 | 7.35523 |
| margin | 353 | 0.197322 | 0.109708 | 0.019915 | 0.641393 |
| tax/revenue | 353 | 0.028583 | 0.020044 | -0.00119 | 0.144987 |
| capiral expenditure | 353 | 146.0977 | 1080.01 | 0.031394 | 19885 |
| retain earnings | 353 | 349.4264 | 3190.827 | -8.85942 | 59817 |
| culture | 353 | 0.767705 | 0.422895 | 0 | 1 |
| industry | 353 | 0.096317 | 0.295445 | 0 | 1 |

The final sample of the year 2007 contains 164 firms in total. Summary statistics are shown in Table 14 panel (b). The final sample of the year 2007 contains 120 firms that IPO in China, 18 firms that IPO in the U.S, 12 firms that IPO in Hongkong and 8 firms that

IPO in Singapore, 2 firms IPO in the U.K. Only one firm that IPO in Japan and two firms IPO in Germany.

From this simple summary, it's not difficult to see that the proportion of the firms that are IT or IT related companies among the U.S IPO companies is higher than that proportion among the China mainland IPO companies. However, the U.S. IPO firms have relatively smaller asset value on average. The U.S IPO firms also have the smallest tax/revenue proportion and smallest capital expenditure on average. However, those firms have the greatest growth rates in revenue in the recent two years than the firms in other groups.

## 3.5 Estimation results and implications

### *3.5.1 Estimation results for the binary response model*

The results of binary response regressions are shown in Table 15.

## Table 15(a) – Logistic regression results of the data of year 2010[142]

**Year 2010 sample**

| | *Logit* | *Logit* | *Logit* |
|---|---|---|---|
| **Variable** | **Coefficient** | **Coefficient** | **Coefficient** |
| ***LEVERAGE*** | 0.0078 | | |
| | (0.744) | | |
| ***ASSET*** | **-0.00336***** | **-0.003416***** | **-0.00336***** |
| | (0.000718) | (0.000586) | (0.000593) |
| ***GROWTH*** | **0.809322***** | **0.758041***** | **0.766***** |
| | (0.2557) | (0.22127) | (0.2203) |
| ***MARGIN*** | -3.212195 | | |
| | (2.8950) | | |
| ***Tax/Rev*** | 0.84193 | | |
| | (15.7910) | | |
| ***CAPEX*** | **0.00481***** | **0.00433***** | **0.00428**** |
| | (0.00162) | (0.00132) | (0.00151) |
| ***D2*** | -0.13391 | | -0.2769 |
| | (0.57831) | | (0.2142) |
| ***D1*** | 0.00781 | | 0.00611 |
| | (0.8561) | | (0.9142) |
| ***INET*ICTAX*** | 0.01087 | 0.00596 | |
| | (0.0.00843) | (0.00748) | |
| ***Equity*GOV*** | **0.00251***** | **0.00232***** | **0.00230***** |
| | (0.000406) | (0.000000183) | (0.000356) |
| ***constant*** | **-2.5527***** | **-3.07312***** | **-3.43430***** |
| | (0.70085) | (0.2731) | (0.31886) |
| ***Log-likelihood*** | -71.8165 | -81.6372 | -76.4526 |
| ***N*** | 409 | 409 | 409 |
| ***Pseudo R2*** | 0.5603 | 0.5002 | 0.536 |

[142] The numbers in the paratheses in this chapter are the MLE standard errors which are estimated by using the information matrix unless otherwise specified.

**Table 15(b) – Logistic regression results of the data of year 2007**

**Year 2007 sample**

| | **Logit** | **Logit** | **Logit** |
|---|---|---|---|
| ***Variable*** | **Coefficient** | **Coefficient** | **Coefficient** |
| ***LEVERAGE*** | -1.019833 | | |
| | (0.7635) | | |
| ***ASSET*** | -0.001293 | | -0.00177 |
| | (0.0013338) | | (0.0014) |
| ***GROWTH*** | 0.5007* | | |
| | (0.2571) | | |
| ***MARGIN*** | 25.007*** | 21.436*** | 24.79061*** |
| | (6.2830) | (4.7462) | (5.7629) |
| ***Tax/Rev*** | **-169.5436*** ** | **-141.4301*** ** | **-162.1941*** ** |
| | **(40.51623)** | **(29.9114)** | **(36.4808)** |
| ***CAPEX*** | 0.0001987 | | 0.000201 |
| | (0.0003853) | | (0.000346) |
| ***D2*** | -1.331796 | | -0.856196 |
| | (1.255679) | | (1.140123) |
| ***D1*** | -0.3527519 | | -0.25921 |
| | (1.324538) | | (1.19777) |
| ***INET*ICTAX*** | **1.360423*** ** | **1.110524*** ** | **1.21941*** ** |
| | **(0.3069814)** | **(0.22488)** | **(0.26225)** |
| ***Equity*GOV*** | 0.0004407 | -0.00000072 | 0.000606 |
| | (0.0004449) | (0.0000071) | (0.0004529) |
| ***constant*** | -4.692221*** | -5.93512*** | -3.95916*** |
| | (1.536001) | (1.02433) | (1.43213) |
| ***Log- likelihood*** | -18.1671 | -56.389 | -20.8550 |
| ***N*** | 164 | 164 | 164 |
| ***Pseudo R2*** | 0.8095 | 0.4088 | 0.7813 |

What we can learn by looking at the results reported in the table is that tax/revenue ratio has no significant effects on decisions of IPO overseas for the 2010 sample but has very significant effects on decisions of IPO overseas for the 2007 sample  If we look at the estimation result for the 2007 sample, we can see some interesting patterns: The coefficient of variable tax/ revenue ratio is always significant and negative. In the meanwhile, the coefficient of the term margin is always significantly positive. That means, the firms paying lower taxes, in other words, having low levels of information disclosure while having high profitability are more willing to IPO overseas. By combing these two coefficients, it's not hard to see that the Chinese firms that pay lower taxes but have higher margins are more willing to IPO overseas. In other words, the Chinese firms that have a lower level of information disclosure are more willing to IPO overseas in 2007. If we look at the coefficient of INET*ICTAX, which measures the expected transparency in the destination market, it's always positively significant. That is to say, the firms that have higher expected transparency in the destination market are more willing to IPO overseas. Combining the coefficients of the term Margin and the term Tax/revenue ratio, it tells us that the incentive for Chinese firms to go public overseas is to improve the firm transparency. The firm's shareholders are sacrificing the high price premium in order to improve firm transparency. This finding perhaps is implying that the Chinese firms that have bad information disclosure, bad corporate governance and server conflicts between managers and shareholders are more likely to go public overseas.

If we look at the estimation results for the sample of the year 2010, there is a structural change in the IPO behaviors. The coefficients of the asset term are always negatively significant. That is somewhat betraying our intuition since based on the previous research by other scholars, larger firms seem to be more intended to IPO or cross-list in other countries. In order to confirm this effect, I also use other proxies for firm size to do the robustness check and I find out that this effect is robust. It means the firms that are smaller are more willing to IPO overseas in 2010. The firms that have higher capital expenditure and higher growth rates are more likely to IPO overseas in 2010, which implies that the firms that develop fast and face borrow constrain are more likely to IPO overseas. Combine this with the coefficient of the size measure (asset value and equity), it's making sense because smaller developing firms usually have lower credit in terms of borrowing. Borrowing constraints will more likely be bounding for these firms. These firms are more likely to raise money by IPO in other countries. The term Equity*GOV is also always positively significant. Basically, it means the quality of the government at the destination market is also a concern when firms are making IPO decisions. Given the two firms are equal in all aspects, the foreign governments that have better quality will make Chinese firms more likely to IPO in foreign countries.

When we look at the difference between the estimation results for the 2010 sample and the 2007 sample, if we compare the Pseudo R-square statistics, we will see that there must be some structural changes during 2008-2009. After the 2008 global financial crisis, all countries, including the Chinese government, are taking the financial market more and more seriously. During that period, the Chinese government tried to make harsher regulations and higher standers for IPO. The risk management system was playing a more important role than before and corporate governance issues became a focus. These changes make the Chinese firm shareholders no longer use oversea IPO as a measure to improve the transparency of the firms that they invest in. Instead, they start to focus more on realistic issues: Development and borrow constrain. After the global financial crisis, firms also pay attention to government quality when they are choosing IPO locations.

### *3.5.2 Multinomial logit regression results*

After doing the IIA test for the suspected alternatives, I find that the IIA property holds for these alternatives. That means by using the multinomial logit regression framework here will still match the random utility framework and will give consistent estimators. Table 16 below shows the results of the MNL regression. For this part, I only looked at the estimation results for the 2010 sample because as discussed above, the preference of firms in terms of IPO changes a lot. Looking at 2010 sample will give us more valid implications. I drop category 4 because there is only one firm in our sample, choosing to IPO in German in 2010. The estimation of that category will be meaningless.

**Table 16 – Multinomial logit regression results of the data of year 2010**

| Year 2010 sample | | | |
|---|---|---|---|
| | Multinomial Logit | | |
| | *IPO in the U.S* | *IPO in Hongkong* | *IPO in Singapore* |
| **Variable** | **Coefficient** | **Coefficient** | **Coefficient** |
| ***LEVERAGE*** | **-0.1806**** | -0.1486 | -0.06189 |
| | (0.08209) | (0.30075) | (0.2749) |
| ***ASSET*** | **-0.005317***** | **-0.00494***** | **-0.0023**** |
| | (0.001042) | (0.00101) | (0.000913) |
| ***GROWTH*** | **1.1064***** | -0.6660 | -0.4372 |
| | (0.2788) | (1.24506) | (0.9831) |
| ***MARGIN*** | -3.7448 | 0.9007 | 0.1554 |
| | (3.503) | (6.32113) | (5.1638) |
| ***Tax/Rev*** | -3.3166 | -7.3063 | -13.5196 |
| | (18.5047) | (30.9740) | (30.4885) |
| ***CAPEX*** | **0.004998*** | **0.007765***** | **-0.01169*** |
| | (0.002772) | (0.00132) | (0.006762) |
| ***D2*** | 0.27206 | -1.753462 | -0.60028 |
| | (0.71824) | (1.2249) | (1.4045) |
| ***D1*** | -0.3727 | 0.2428 | 1.2586 |
| | (0.96685) | (1.7702) | (1.5347) |
| ***INET*ICTAX*** | 0.01434 | 0.0000221 | **0.16544*** |
| | (0.009417) | (0.013792) | (0.08573) |
| ***Equity*GOV*** | **0.0031834***** | **0.003144***** | **0.002186***** |
| | (0.000523) | (0.0005461) | (0.000586) |
| ***constant*** | **-2.6336***** | **-4.06625** | **-5.5183***** |
| | (0.90833) | (1.39192) | (1.7451) |
| ***Base Category*** | IPO in China Mainland | | |
| ***log likelihood*** | -90.2526 | | |
| ***N*** | 408 | | |
| ***Pseudo R2*** | 0.5745 | | |

From the table of estimation results, it's not difficult to learn that, after the financial crisis, information disclosure and firm transparency is no longer a key concern when Chinese firms are choosing IPO locations. Government quality becomes a huge concern in their decision making. No matter where they IPO, they always care about government quality in the destination country. By fixing everything else, if the government quality is better, then the odds ratio of IPO in the U.S., IPO in Hongkong and IPO in Singapore relative to IPO in China mainland will be significantly higher.

If we look at the estimation results at the alternative level, it’s evident that the odds ratio between IPO in China mainland and IPO in the U.S, is also significantly influenced by the firm size. Again, it shows that smaller firms have higher odds of IPO in the U.S relative to IPO in China Mainland. In the meanwhile, the firms that grow faster and have more capital expenditure also are more likely to IPO in the U.S. Firms that are smaller and have higher capital expenditure are also in favor of IPO in Hongkong. Those findings may imply that smaller and less developed firms that are facing borrowing constraints are more willing to IPO in the U.S. and in Hongkong. Among those firms, the firms that grow fast will more prefer IPO in the U.S.

To sum up, based on our analysis from the perspective of discrete choice modeling, I find that the firm size, borrowing constraints and government quality at the destination market are the key determinants that drive Chinese firms' IPO overseas in 2010. The culture effect is not significant when explaining the decision of IPO in the U.S. There’s no evidence showing firms in the IT industry are more willing to IPO overseas. The growth rate is also a key concern especially for the firms that want to IPO in the U.S.

### 3.6 Ex-post analysis

Though the ex-ante analysis can show the motivations of the Chinese firms to IPO overseas to a great extent, it doesn’t tell the whole story. Especially, the ex-ante analysis only focuses on two specific years and it only shows the ex-ante motivations of those firms who IPO during those periods. Blass et al (2001) also do an ex-post analysis in addition to the ex-ante analysis in their research which aims to explore the answer to the question of why firms go public. In that research, the purpose of the ex-post analysis is to show what actually changes after the firms go public and those actual changes may become the reasons for other firms to go public. The logic in this paper is quite similar in the sense that: Though there are some factors that firms had considered about before they made the IPO, there are some other factors that those firms hadn't considered about. Significant differences in those factors between the firms which IPO locally and IPO overseas may become the reasons for other firms to IPO oversea or IPO locally. Therefore, I am also conducting another ex-post analysis.

The key methodology that I use in this analysis is the Diff-in-diff method. The main specification for the Diff-in-diff analysis is

$$\mathrm{y}_{it} = \beta_0 + \beta_1 afterIPO_{it} + \beta_2 IPOoverseas_i + \beta_3 afterIPO_{it} * IPOoverseas_i + \delta X_{it} + \vartheta T + \vartheta I + \varepsilon_{it} \tag{11}$$

Where $\mathrm{y}_{it}$ is a dependent variable that includes performance indicators that are possibly different between firms that IPO locally and firms that IPO overseas. The $afterIPO_{it}$ the term is the dummy variable which equals to 1 if firm i has already gone public at year t, $IPOoverseas_i$ is the dummy variable which equals 1 if firm i IPO overseas. $X_{it}$ is the vector of firm-level controls. T is the time fixed effect and I is the industry fixed effect.

Since we know that the decision of IPO oversea is determined by some firm-level factors, variable $IPOoverseas_i$ is essentially endogenous. In order to resolve this kind of endogenous problem, since we already know that growth, capital expenditure, firm size, information disclosure level and equity scale, financial leverages can be the factors that determine the IPO location, we need to control all these factors and include all of them in the $X_{it}$ vector in order to minimize the bias caused by self-selection.

The independent variables that I am interested here include Return on Assets (ROA), Return on Equity (ROE), profit margins, current ratios, the portion of operating income and the portion of non-operational income. The reason why I am looking at these variables is that the variables mentioned above are good measurements for performances, risk management and the quality of incomes. These are some long-term factors that firms may want to consider when making IPO choices.

The data that I am using in this analysis is also from RESSET database and the CSMAR database. The final sample made up of 6822 firms and 37492 firm-year combinations. 810 firms IPO overseas and 6012 firms IPO locally. The summary statistics are attached in Appendix A table 2. As for the estimation, I use OLS estimation with robust standard error. The results are shown below:

**Table 17 – Difference- in- difference estimation results for the Ex- post analysis**

| Diff-in-Diff estimation | | | |
|---|---|---|---|
| ***Variables*** | **ROA** | **ROE** | **Profit Margin** |
| ***IPO*** | -8.57***<br>(0.35) | -0.13***<br>(0.02) | 0.37<br>(0.31) |
| ***IPOoverseas*** | -28.05***<br>(1.74) | -0.31**<br>(0.16) | -2.06<br>(1.42) |
| ***IPO*IPOoverseas*** | 9.12***<br>(0.26) | 0.016<br>(0.06) | -1.38*<br>(0.71) |
| ***Controls*** | Yes | Yes | Yes |
| ***Year fixed-effectss*** | Yes | Yes | Yes |
| ***Industry fixed-effectss*** | Yes | Yes | Yes |
| N | 24,325 | 24,330 | 24,330 |
| adj.R-sq | 0.1931 | 0.4652 | 0.3736 |
| F-stat | 2444.59 | 239.27 | 2.72 |

| ***Variables*** | **Current ratio** | **Oprational income to total profit** |
|---|---|---|
| ***IPO*** | 1.63***<br>(0.17) | -0.014<br>(0.017) |
| ***IPOoverseas*** | -7.84***<br>(0.84) | 1.16***<br>(0.24) |
| ***IPO*IPOoverseas*** | -0.89***<br>(0.24) | -0.34***<br>(0.15) |
| ***Controls*** | Yes | Yes |
| ***Year fixed-effectss*** | Yes | Yes |
| ***Industry fixed-effectss*** | Yes | Yes |
| N | 24,152 | 24,320 |
| adj.R-sq | 0.0285 | 0.0321 |
| F-stat | 55.66 | 49642.53 |

From the results shown above in Table 4 (b), the real effects from oversea IPOs on firm profitability are positive though the estimator in the regression on ROE is not significant. Essentially, the estimation results show that the firms are doing worse in terms of profitability after IPO and the firms which make IPOs overseas suffer even more over the whole sample period. However, the firms that make IPOs overseas are doing better once they finish the IPO than the firms that make IPOs locally.

The profit margin is a measure to see what the proportion of expense is relative to the whole revenue for a specific firm. According to the regression results, the real effects of oversea IPOs on firm profit margins are significantly negative. However, as we can see from the estimation results of regressions on profitability (ROA and ROE) that the effects of oversea IPOs on firm profitability are positive. As we know ROA=profit margin*asset turnover, then we can also infer that the effects from oversea IPOs on firm asset turnovers are positive.

By applying the same logic of interpretation, the real effects from oversea IPOs on the current ratio are negative, which implies that the firms that make IPOs overseas will be less liquidate and have higher default risks. Basically, it implies that the firm's IPO overseas becomes riskier. The reason for this phenomenon is not clear but this definitely can be something interesting to conduct research on.

If we look at the last two columns of table 4, the two columns are telling us that the firms that IPO overseas have different income structures from the firm's IPO locally. The firms that IPO overseas have more incomes from the non-operation side: investments, sales of properties, currency arbitrages and non-recursive gains.

## 3.7 Robustness check

### *3.7.1 Sample size and bootstrapping*

One of the concerns is that our sample sizes are somewhat small in both the 2007 sample and 2010 samples. This is the limitation of the data: some data for the firms that IPO overseas is missing, leaving us with a relatively small sample that can use for analysis. In order to overcome that issue a little bit, I perform a robustness check by using the bootstrapping strategy.

The bootstrapping strategy is leveraging the resampling strategy to estimate the distribution of the parameters and then we can make inference based on the expectation value of the distribution. I run 500 times of repetition for both samples. The results still remain robust, and my finds remain valid though the significance of the parameters does change a little bit. The results are attached in the Appendix C.1.

### *3.7.2 Measurements of firm size*

According to the research conducted by Dang et al. (2015), the market value, total asset value, and market caps are three robust measurements for firm size, which will induce the least measurement error issue. However, I think it's still necessary by replacing total asset value with other possible proxies to test the robustness of the size effect and the empirical specification.

Therefore, I also gather the data of market value and market caps from RESSET database. Using these data, I replicate the regression, and the results remain the same though the magnitudes of the estimators change. The interpretation and conclusion regarding firm size effect on the choice of IPO remain the same and robust. Due to the limitation of the space, the results are not displayed.

### *3.7.3 Subsample regression*

Since there is no specific industrial effect that has been found in the estimation section, a subsample generated based on the industry should also follow the patterns that we have already found. Thus, in this part, I do a subsample regression by dropping all the firms in the IT industry. The subsample regression results confirm the robustness of my findings. The results are attached in Appendix C.1.

### *3.7.4 Probit model*

As mentioned earlier, since I don't know whether the error term of the utility function is following a standard normal or a logistic distribution, I need to run regressions on both models and see which model has a better fitting of data. As shown in the appendix, in the binary case, for the 2010 sample, the logit model fits slightly better. For the 2007 sample, if we look at the log-likelihood ratio and the Pseudo R2 statistics, the logistic model fits the data a little bit better. The coefficients' significance doesn't change and signs don't change either. Our conclusion remains robust if we look at the probit models. The results for the probit regressions are attached in Appendix C.1 as well.

### *3.7.5 Retained earnings*

As Kaplan and Luigi (1997) point out in their paper, the firms can finance their expenditure by using internal cash flows. Internal cash flow is even cheaper in terms of cost. If a firm has large retain earnings, they may not even want to use common stock financing. The best measure for total internal cash flow is retained earnings. In order to make sure our result without retain earning is still robust, I add in the variable of retained earnings into different models and previous findings still remain robust. The coefficient of the retained earning variable is not significant at all, which means the retained earning level is not a key concern that the firms care about when they are making IPO location decisions. The results are reported in Appendix C.1.

### *3.7.6 Alternative proxies for government quality*

In order to make sure our conclusion about the 2010 sample is correct, I use several other types of proxies for government quality. As mentioned, there are six indexes for the government quality included in the Worldwide Governance Indicators database, which are Voice and Accountability indicator, Political Stability and Absence of Violence/Terrorism indicator, Government Effectiveness indicator, Regulatory Quality indicator, Rule of Law Control of Corruption indicators. In this subsection, I use the Political Stability and Absence of Violence/Terrorism indicator, Regulatory Quality indicator, and Control of Corruption indicator to proxy for government quality.

The results are attached in the Appendix C.1. Basically, no matter which indicators I use to proxy for government quality, the results don’t change. There are very slight changes in the significance of the coefficient of capital expenditure but it’s still significant at the 95% confidence level. If we look at the log-likelihood and Pseudo R-square numbers reported in the estimation results, the Corruption Control indicator gives the highest log-likelihood and Pseudo R-square compared to the other two indicators. It seems that corruption control is a key consideration when they are deciding IPO locations after 2010.

## 3.8 Conclusion

In this paper, I try to explore what determinants make Chinese firms IPO overseas, giving up the huge premium offered by the domestic investors in the market of China mainland. Based on previous literature regarding similar research, I build up empirical models to test which determinants are the key determinants that make Chinese firms IPO overseas. Also, I decompose the choice of IPO overseas for the 2010 sample and explore the determinants that make the firms IPO in the U.S. and other countries/areas.

The key finding of this paper is that in fact, huge structural changes have been observed in Chinese firms' preferences over IPO locations. The firm information disclosure and firm transparency level after IPO, which are indicated by the combination of tax/revenue ratio, profit margin and the interaction term of internet user ratio multiplied by the income tax ratio, are key and robust determinants when firms making IPO choices in the year 2007. The estimation results also imply that firms which have bad information disclosure and shareholder protection are more likely to IPO overseas. It seems telling us the shareholders of those firms are willing to give up the huge price premium offered by the local investor for the sake of firm transparency. This finding is sort of supporting the findings of previous research.

From the analysis for the firms that IPO in 2010, I observe that small firms with higher growth rate and facing the borrowing constrains are more willing to IPO overseas, especially in the U.S. Also, firms do consider the quality of the government in the destination market when choosing where to IPO. It seems that firms care more about corruption control in the destination market when they try to make IPO decisions.

In addition to the ex-ante analysis, I also conducted an ex-post analysis to find out what may become the reasons for firms to IPO overseas or to IPO locally. The ex-post analysis shows that the firms that IPO overseas are doing better in making profits, are having lower profit margin but higher asset turnover speeds, are having more default risks and are having more incomes from the non-operation side. By looking at these results, other firms may make their IPO decisions accordingly.

Though these findings are proved to be robust by the robustness checks, there are many imperfections and limitations. Though the sample is relatively comprehensive and close to the population, some key information is missing. For instance, information regarding R&D is not available. The detailed corporate governance information is not available. The reason for this is complicated and involves social and political issues. If, in the future, a more comprehensive and complete dataset is available, we can find out more determinants influencing the IPO behaviors of firms.

There are other imperfections in addition to the problem of the data: Endogenous repressors. The firm-level characteristics and the firm's decision can possibly determine simultaneously in the sense that firms can make an IPO decision first and then they start to adjust their financial reports based on the requirement of the destination market. In that sense, the firm-level characteristics that can be obtained from financial reports can be endogenous. This issue can be overcome by using instrumental variables. However, it's also very difficult to find the perfect instrumental variables because one can always argue that this variable is adjusted by firms based on their IPO choices.

Though there are some imperfections in this paper, I believe this paper has contributed to the corporate finance literature regarding the IPO behaviors of Chinese firms. The Chinese equity market is emerging in a quick but peculiar way. Learning more information regarding firms' IPO behaviors can help investors to make better decisions. Also, my research in this paper gives some policy implications about the structural change in the Chinese firms' IPO behaviors after the global financial crisis.

# APPENDIX A. ROBUSTNESS CHECKS AND FIGURES OF CHAPTER ONE

## A.1 The fitting of linear functional form and polynomial functional form

**Table 1 – Linear Projection v.s. Polynomial**

***Linear Projection v.s. Polynomial***

| | *Dependent variable:* | |
|---|---|---|
| | *ROA* | *ROA* |
| | (1) | (2) |
| | ***OLS*** | ***Cubic Polynomial*** |
| *Hausman-type IV* | 7.68***<br>(1.52) | **37.90*****<br>**(6.99)** |
| *Hausman-type IV -sq* | | -529.79***<br>(153.99) |
| *Hausman-type IV -cubic* | | 2171.37***<br>(781.98) |
| *Constant* | 5.31***<br>(0.15) | 5.37***<br>(0.19) |
| ***Adjusted R-sq*** | **0.0018** | **0.0037** |
| *N* | 13501 | 13501 |
| *Industry fixed effect* | No | No |
| *Year fixed effect* | No | NO |
| *Province fixed effect* | No | NO |

## A.2 Description of the text mining process

In particular, I wrote a Python Script to text mine from the annual report in order to collect gain more information on employee education. The Application Developer Interface (API) on the website of the Shanghai Stock Exchange (SSE) follows a simple logic. If we add the date[143], the stock code assigned by SSE and the year as well as the type of the report (n for annual report) at the end of the http address "http://www.sse.com.cn/disclosure/listedinfo/announcement/c", for instance, we can get something like http://www.sse.com.cn/disclosure/listedinfo/announcement/c/2017-04-01/600000_2016_n.pdf. By using the "request" function in urlutil3 module in Python, we can download the pdf files into a folder . Then, by using loops with the OS module[144], text mining algorithm[145] can be conducted on each pdf file and key information can be extracted and put into a data frame.

The key words that are used in our algorithm to scrape the total number employees include "雇员总数"，"员工总数"，"员工数量"，"员工人数"，"员工共"[146]。 The key words in our algorithm to abstract the total number of employees with bachelor degrees include, "本科学历共"，"本科学历员工"，"本科及以上学历员工", "本科及以上学历员工总数"[147]。After cleaning the data, I merge the data to the firm level data obtained from RESSET by using the stock code.

## A.3 Marginal effects of the OLS estimation with different control functions

[143] Usually annual report submitting dates for Chinese public firms include the last day and the first day of each month.

[144] The OS module is a module in standard Python library to help with the interactions between Python developing environment and the operating system. In order to loop through a directory, OS module is essential.

[145] The PDFMiner module in Python is used for the purpose of conversion from PDF to HTML text. The NTLK toolbox is used for natural language processing and conversion from HTML to text token.

[146] All those Chinese phrases have the same meaning of "the total number of employees".

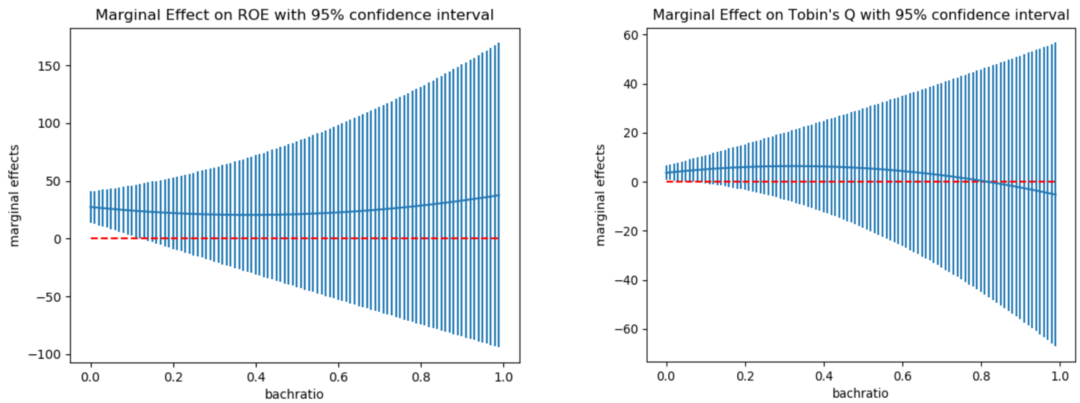


**Figure 1 – Marginal effects on alternative firm performance measurements.**[148]

## A.4 Stock performance robustness checks

### *A.4.1 Clustering of firms*

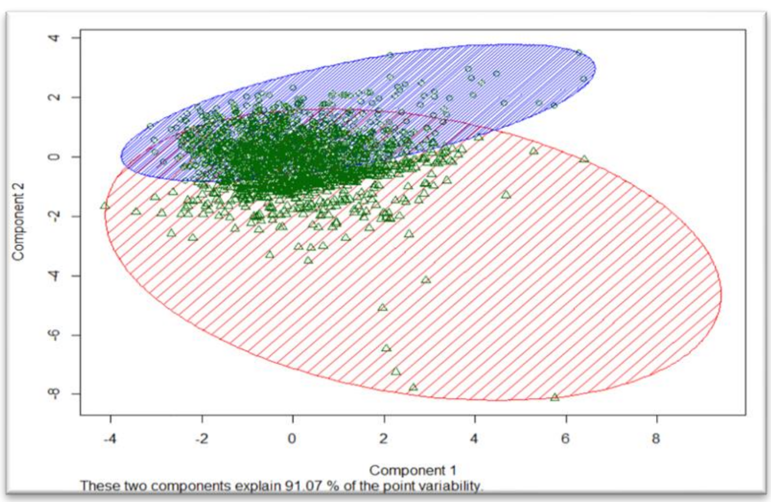


**Figure 2 – Clustering results from MNM clustering algorithm.**[149]

---

[148] The upper panel is based on the estimation results of OLS with MARS control function approach. The lower panel is based on the estimation results of OLS with linear control function approach.

[149] The clustering is based on the asset value which measures the capital level of the firm, the number of employees which measures the labor level of the firm and age. The clustering results show that 91.07% of the cross-group variation has been explained by those variables. 3, The components are computed through the linear combinations of the 3 variables

*A.4.2 Non-parametric tests on CARs*

**Table 3 – Non-parametric Tests on CARs**[150]

***Non-parametric Comparison***

***Panel A. Above Average v.s. Below Average***

| Group | bachratio>=0.24 | bachratio<0.24 |
|---|---|---|
| Avg. CAR | 0.206 | 0.1722 |
| std. error | 0.0099 | 0.0073 |
| N | 6904 | 3918 |

| Hypothesis | ***H0:CAR(bachratio>=0.24) > CAR(bachratio<0.24)*** |
|---|---|
| t-stats | 2.7436 |
| Pr(T>t) | 0.003 |

***Panel B. Below Critical Value v.s. Above Critical Value(Cluster 1)***

| Group | bachratio>=0.1312 | bachratio<0.1312 |
|---|---|---|
| Avg. CAR | 0.0302 | 0.089 |
| std. error | 0.0142 | 0.0206 |
| N | 726 | 217 |

| Hypothesis | ***H0:CAR(bachratio<=0.1312)>CAR(bachratio>0.1312)*** |
|---|---|
| t-stats | 2.0728 |
| Pr(T>t) | 0.019 |

***Panel C. Below Critical Value v.s. Above Critical Value(Cluster 2)***

| Group | bachratio>=0.1312 | bachratio<0.1312 |
|---|---|---|
| Avg. CAR | 0.0086 | 0.116 |
| std. error | 0.0084 | 0.039 |
| N | 584 | 161 |

| Hypothesis | ***H0:CAR(bachratio<=0.1312)>CAR(bachratio>0.1312)*** |
|---|---|
| t-stats | 4.206 |
| Pr(T>t) | 0 |

---

in order to show the clustering results on 2 dimensions. 4, Circle means the data point belong to cluster 1 while triangle means the data point belong to cluster 2.

[150] Table 3 shows the Non-parametric test results on CARs of different groups. Panel A shows the test result of ***bachratio*** above average against below average group for the whole sample. Panel B shows the test results of the first cluster with ***bachratio >0.1312*** group against ***bachratio <=0.1312*** group. Panel C shows the test results of the second cluster with ***bachratio >0.1312*** group against ***bachratio <=0.1312*** group. The threshold value 0.1312 is from the baseline results since the firms have ***bachratio*** larger than 0.1312 is less likely to experience improvement of firm performance.

### A.5 Marginal effects of the fixed-effects estimator

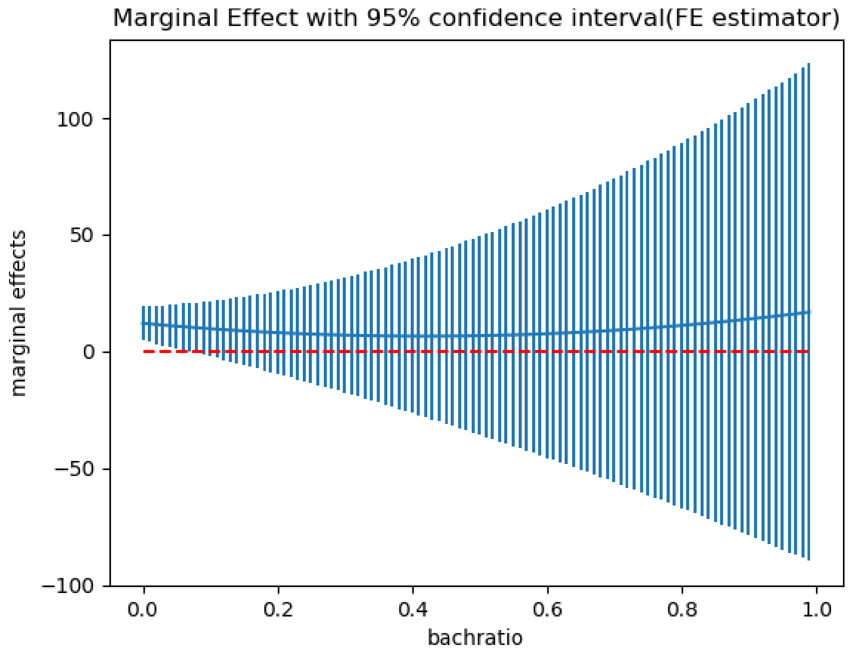


**Figure 4 – Marginal effects with 95% confidence interval**

## A.6 Robustness checks of using different measurements of firm performance

**Table 2 – Robustness checks with alternative measurements** [151]

| | ***Alternative Measurements for Firm performance*** | | | |
|---|---|---|---|---|
| | *Dependent variable:* | | | |
| | *ROE* | | *Tobin's Q* | |
| | (1) | (2) | (3) | (4) |
| | ***OLS with MARS control function approach*** | ***OLS with linear control function approach*** | ***OLS with MARS control function approach*** | ***OLS with linaer control function approach*** |
| *bachratio* | **27.44*** (3.44)** | **43.95*** (9.04)** | **2.54*** (0.73)** | **3.67*** (0.64)** |
| *bachratio-sq* | -17.92* (10.30) | -24.68* (13.53) | 7.64*** (2.96) | 8.48*** (3.20) |
| *bachratio-cubic* | 15.54* (9.23) | 20.89* (12.02) | -8.29*** (2.67) | -8.74*** (2.87) |
| *Log(asset)* | -9.12*** (0.22) | -9.04*** (0.24) | -1.26*** (0.08) | -0.46*** (0.09) |
| *Log(emp)* | 0.83*** (0.20) | 1.82*** (0.42) | 0.40*** (0.04) | -0.46*** (0.09) |
| *Log(Avg_Salary)* | **-0.21** (0.076)** | **-0.41*** (0.17)** | **0.05*** (0.016)** | **0.27*** (0.02)** |
| *age* | -0.062*** (0.02) | -0.073*** (0.01) | -0.17*** (0.004) | 0.018*** (0.003) |
| *Log(exp)* | 8.71*** (0.24) | 8.09*** (0.25) | 0.21*** (0.08) | 0.488*** (0.09) |
| *markup* | 69.24*** (1.35) | 64.49*** (1.93) | 5.39*** (0.29) | 8.003*** (0.41) |
| *Log(Q)* | 0.15*** (0.04) | 0.36*** (0.10) | 0.14*** (0.01) | -0.034** (0.017) |
| $\hat{v}$ | -22.32*** (1.42) | -33.51*** (7.05) | -1.92*** (0.53) | 11.59*** (0.) |
| *Constant* | 4.71*** (2.08) | 3.12* (2.20) | 16.30*** (0.53) | 3.12* (2.20) |
| *Adjusted R-sq* | 0.27 | 0.26 | 0.29 | 0.30 |
| *N* | 13481 | 13481 | 13501 | 13501 |
| *Industry fixed effect* | Yes | Yes | Yes | Yes |
| *Year fixed effect* | Yes | Yes | Yes | Yes |
| *Province fixed effect* | Yes | Yes | Yes | Yes |

[151] The numbers in the parentheses are the bootstrapping standard errors. The variable ***bachratio*** *is the aggregated education level of employees measured by the proportion of proportion of employees holding at least bachelor's degree or above. Log(exp) is the natural log transformation of expenditure. Log(rev) is the* $\hat{v}$ is the residual generated by using the control function approach.

## A.7 Proofs

*A.7.1 Proof for the net effect of aggregated employee education level on firm performance*

Since we have the presumed model: $\boldsymbol{y_{it} = a + f(bachratio_{it}) + \delta X_{it} + \theta D_{it} + \varepsilon_{it}}$ ,

Then $\boldsymbol{E[y_{it}|bachratio_{it}, X_{it}, D_{it}] = a + f(bachratio_{it}) + \delta X_{it} + \theta D_{it}}$.

Based on the property of conditional expectation:

$$\boldsymbol{E[y_{it}|bachratio_{it}]}$$
$$\boldsymbol{= a + f(bachratio_{it}) + \delta E[X_{it}|bachratio_{it}]}$$
$$\boldsymbol{+ \theta E[D_{it}|bachratio_{it}]}$$

Since assumed that $\boldsymbol{X_{it}}$ and $\boldsymbol{D_{it}}$ are exogenous, $\boldsymbol{X_{it}}$ are approximately seen as independent of $\boldsymbol{bachratio_{it}}$. Also, the location and year fixed effects are not influencing or influenced by the employee's education level of firm it, $\boldsymbol{D_{it}}$ are independent of $\boldsymbol{bachratio_{it}}$.

Based on the Bayes' rule,

$$\boldsymbol{E[y_{it}|bachratio_{it}]}$$
$$\boldsymbol{= a + f(bachratio_{it}) + \delta E[X_{it}|bachratio_{it}]}$$
$$\boldsymbol{+ \theta E[D_{it}|bachratio_{it}] = a + f(bachratio_{it}) + \delta E[X_{it}] + \theta E[D_{it}]}$$

$$\because \boldsymbol{y_{it} = E[y_{it}|bachratio_{it}, X_{it}, D_{it}] + \varepsilon_{it} = E[y_{it}|bachratio_{it}] + \varepsilon_{it}}$$
$$\boldsymbol{= a + f(bachratio_{it}) + \delta E[X_{it}] + \theta E[D_{it}] + \varepsilon_{it}}$$

Based on the linear property,

$$\therefore \boldsymbol{y_{it} - a - \delta E[X_{it}] + \theta E[D_{it}] = f(bachratio_{it}) + \varepsilon_{it}}$$

*A.7.2 Proof for the inability of IV estimator to identity marginal effects when the relationship between the IV and the dependent variable is not linear.*

Having linear relationship between the endogenous variable x and the dependent variable, we have:

$$\boldsymbol{y = a + b_1 x + \mathrm{e}}$$

For simplicity, assuming only one independent variable and only one endogenous variable existing in the model, the conclusion can still be generalized to multi-variables regression cases.

When having linear structure between the dependent variable and the instrumental variable z, we have

$$\boldsymbol{\frac{dy}{dz} = \frac{d(y = a_1 + \theta z + \epsilon_1)}{dz} = \theta = \frac{cov(y, z)}{var(z)}}$$

In matrix form we can write it as $(\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'Y}$.

When having linear structure between the independent variable and the instrumental variable, we have

$$\boldsymbol{\frac{dx}{dz} = \frac{d(x = \alpha_2 + \rho z + \epsilon_2)}{dz} = \rho = \frac{cov(x, z)}{var(z)}}$$

In matrix form we can write it as $(\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'X}$.

Therefore, the true marginal effect from x on y is

$$\frac{dy}{dx}=\frac{\frac{dy}{dz}}{\frac{dx}{dz}}=\frac{\theta}{\rho}=\frac{\frac{cov(y,z)}{var(z)}}{\frac{cov(x,z)}{var(z)}}=\frac{cov(y,z)}{cov(x,z)}=\boldsymbol{b_1}$$

By written in matrix form we have $(\boldsymbol{Z'X})^{-1}(\boldsymbol{Z'Y})$, which is the IV estimator $\boldsymbol{b}_{1,IV}$.

When having non-linear relationship between the dependent variable and the instrumental variable, then

$$\frac{dy}{dz}=\frac{d(y(z))}{dz}\neq\frac{d(y=a_1+\theta z+\epsilon_1)}{dz}=\theta=\frac{cov(y,z)}{var(z)}$$

Then $\boldsymbol{b_1}=\frac{dy}{dx}=\frac{\frac{dy}{dz}}{\frac{dx}{dz}}=\frac{\frac{d(y(z))}{dz}}{\frac{dx}{dz}}\neq\frac{\frac{d(y=a_1+\theta z+\epsilon_1)}{dz}}{\frac{dx}{dz}}=\frac{cov(y,z)}{cov(x,z)}=(\boldsymbol{Z'X})^{-1}(\boldsymbol{Z'Y})=\boldsymbol{b}_{1,IV}$

Therefore, $\frac{\frac{dy}{dz}}{\frac{dx}{dz}}$can no longer be used to identify the coefficient $\boldsymbol{b_1}$.

*A.7.3 Calculation of the standard error of the marginal effect*

Generally, assume a stochastic cubic polynomial function $\boldsymbol{m}=\boldsymbol{ax}+\boldsymbol{bx^2}+\boldsymbol{cx^3}$, where $\boldsymbol{a},\boldsymbol{b}\ \boldsymbol{and}\ \boldsymbol{c}$ are random variables which follow normal distribution and the variance-covariance matrix for $\boldsymbol{a},\boldsymbol{b}\ \boldsymbol{and}\ \boldsymbol{c}$ is defined to be *V*. *V* is 3X3 matrix.

$$V=\begin{bmatrix}\mathbf{var(a)} & \mathbf{cov(a,b)} & \mathbf{cov(a,c)}\\ \mathbf{cov(b,a)} & \mathbf{var(b)} & \mathbf{cov(b,c)}\\ \mathbf{cov(c,a)} & \mathbf{cov(c,b)} & \mathbf{var(c)}\end{bmatrix}$$

The marginal effect from $\boldsymbol{x}$ on $\boldsymbol{m}$ can be obtained from the first order derivative:$\frac{\partial m}{\partial x}=\boldsymbol{a}+\boldsymbol{2bx}+\boldsymbol{3cx^2}$.

Based on the rule of variance, $\boldsymbol{var}\left(\frac{\partial m}{\partial x}\right)=\boldsymbol{var}(\boldsymbol{a}+\boldsymbol{2bx}+\boldsymbol{3cx^2})=\Omega$, where $\Omega=\begin{bmatrix} \mathbf{var}(\mathbf{a}) & \mathbf{cov}(\mathbf{a}, 2\mathbf{b}\boldsymbol{x}) & \mathbf{cov}(\mathbf{a}, 3\mathbf{c}\boldsymbol{x}^2) \\ \mathbf{cov}(2\mathbf{b}\boldsymbol{x}, \mathbf{a}) & \mathbf{var}(2\mathbf{b}\boldsymbol{x}) & \mathbf{cov}(2\mathbf{b}\boldsymbol{x}, 3\mathbf{c}\boldsymbol{x}^2) \\ \mathbf{cov}(3\mathbf{c}\boldsymbol{x}^2, \mathbf{a}) & \mathbf{cov}(3\mathbf{c}\boldsymbol{x}^2, 2\mathbf{b}\boldsymbol{x}) & \mathbf{var}(3\mathbf{c}\boldsymbol{x}^2) \end{bmatrix}$

Based on the rule of covariance and rule of variance:

$$\Omega=\begin{bmatrix} \mathbf{var}(\mathbf{a}) & 2\boldsymbol{x} * \mathbf{cov}(\mathbf{a}, \mathbf{b}) & 3\boldsymbol{x}^2 * \mathbf{cov}(\mathbf{a}, \mathbf{c}) \\ 2\boldsymbol{x} * \mathbf{cov}(\mathbf{b}, \mathbf{a}) & 4\boldsymbol{x}^2 * \mathbf{var}(\mathbf{b}) & 6\boldsymbol{x}^3 * \mathbf{cov}(\mathbf{b}, \mathbf{c}) \\ 9\boldsymbol{x}^4 * \mathbf{cov}(\mathbf{c}, \mathbf{a}) & 6\boldsymbol{x}^3 * \mathbf{cov}(\mathbf{c}, \mathbf{b}) & 9\boldsymbol{x}^4 * \mathbf{var}(\mathbf{c}) \end{bmatrix}$$

Since by doing OLS estimation, the estimated asymptotic variance matrix $\widehat{\boldsymbol{V}}$ can be obtained, $\Omega$ can be obtained by plugging in $\boldsymbol{x}$. $\Omega^{\mathbf{1/2}}$ will be the standard error for the marginal effect from $\boldsymbol{x}_{|x=x_0}$, where $\boldsymbol{x_0}$ is the current value of $\boldsymbol{x}$.

Since $\boldsymbol{a}, \boldsymbol{b}$ $\boldsymbol{and}$ $\boldsymbol{c}$ follow normal distribution, $\frac{\partial \boldsymbol{m}}{\partial \boldsymbol{x}}$ follow normal distribution based upon the property of random distribution.

$$\frac{\partial \boldsymbol{m}}{\partial \boldsymbol{x}}_{|x=x_0} \sim \boldsymbol{N}(\boldsymbol{E}[\boldsymbol{a}] + 2\boldsymbol{x_0}\boldsymbol{E}[\boldsymbol{b}] + 3\boldsymbol{x_0}^2\boldsymbol{E}[\boldsymbol{c}], \Omega_{|x=x_0})$$

The confidence interval can be computed accordingly.

# APPENDIX B. ROBUSTNESS CHECKS AND ADDITIONAL DATA OF CHAPTER TWO

## B.1 Robustness checks

### Table 1 – Robustness checks with the control function approach[152]

***Control function approach with BLP framework***

| | Dependent variable: $\log(share_{it}) - \log(share_{0t})$ | |
|---|---|---|
| | (1) | (2) |
| | ***Control function OLS on whole sample*** | ***Control function OLS on subsample*** |
| Loan interest rate | **-100.91*****<br>**(10.80)** | **-71.47*****<br>**(16.19)** |
| Number of employees | 0.00015***<br>(8.13e-06) | 0.000066***<br>(0.00014) |
| Goodwill to total asset ratio | 6.48<br>(0.75) | 26.42***<br>(5.92) |
| Intangible asset per employee | 0.11<br>(0.22) | 20.66***<br>(4.79) |
| Liquid asset to total asset ratio | -0.005***<br>(0.002) | -0.06***<br>(0.02) |
| Asset per employee | 9.81e-06***<br>(3.00e-06) | -2.34e-06<br>(9.49e-06) |
| Average staff expenses | -0.0002<br>(0.0003) | 0.02***<br>(0.01) |
| Total asset | - 1.12e-08<br>(1.00e-09) | -3.05e-09*<br>(1.69e-09) |
| $\hat{v}$ | 123.2409***<br>(10.88011) | 93.98***<br>(18.4) |
| constant | -2.85***<br>(0.54) | -8.19***<br>(1.24) |
| *Adjusted R-sq* | 0.25 | 0.63 |
| *N* | 5476 | 240 |
| F (9,N-10) | 200.71 | 46.42 |

[152] Table 1 shows the results of control function approach estimation combined with the BLP framework. The estimation specification still follows the structural model framework proposed in Berry (1994). However, since the random utility from the price factor is not significant, the simulation technique is not used here and assume the individuals all have the same preference over interest rate. The numbers in the parentheses are the standard errors. $\hat{v}$ denotes the control of the endogeneity.

**Table 2 – Robustness checks on the size effects**[153]

| ***Testing Size effects*** | | |
|---|---|---|
| | *Dependent variable:* $\log(share_{it}) - \log(share_{0t})$ | |
| | (1) | (2) |
| | ***Using the actual asset value in the interaction term on the whole sample*** | ***Using the size dummy in the interaction term on the whole sample*** |
| Loan interest rate | **-138.16*** (11.044)** | **-126.83*** (9.13)** |
| Number of employees | 0.001*** (0.00002) | 0.000055*** (0.00001) |
| Goodwill to total asset ratio | 6.38 (0.75) | 2.13*** (0.67) |
| Intangible asset per employee | 0.18 (0.22) | 0.10 (0.32) |
| Liquid asset to total asset ratio | -0.006** (0.003) | -0.008* (0.004) |
| Asset per employee | 6.93e-06** (3.09e-06) | 0.000002 (9.49e-06) |
| Average staff expenses | 0.00008 (0.0003) | 0.0007 (0.0006) |
| Total asset | - 9.93e-06 (1.07e-06) | -3.05e-06*** (1.69e-07) |
| Size * Loan interest rate | 0.00013*** (0.00003) | 68.90*** (1.01) |
| constant | -1.33** (0.55) | -2.04*** (0.46) |
| *Adjusted R-sq* | 0.24 | 0.53 |
| *N* | 5476 | 5476 |
| F (9,N-10) | 193.65 | 1126.06 |

[153] Table 2 shows the results of the estimation by adding an interaction term between the size, measured by total book values of assets, and loan interest rates. The estimation specification still follows the structural model framework proposed in Berry (1994). However, since the random utility from the price factor is not significant, the simulation technique is not used here and assumes the individuals all have the same preference over the interest rate. The numbers in the parentheses are the standard errors. When the book value of assets is larger than the average, the dummy equals to 1. Otherwise, the dummy equals to 0.

**Table 3(a) – Robustness checks with random sampling**[154]

| | ***BLP Estimation*** | |
|---|---|---|
| | *Dependent variable:* $\log(share_{it}) - \log(share_{0t})$ | |
| | (1) | |
| | ***BLP estimation*** | $\sum$ |
| Loan interest rate | **-161.70*****<br>**(87.55)** | **32.37**<br>**(56.67)** |
| Number of employees | 0.00013***<br>(0.000036) | |
| Goodwill to total asset ratio | 0.00013***<br>(0.00005) | |
| Intangible asset per employee | -0.022<br>(0.023) | |
| Liquid asset to total asset ratio | -0.017<br>(0.03) | |
| Asset per employee | 0.00002<br>(0.000024) | |
| Average staff expenses | 0.0015**<br>(0.0002) | |
| Total asset | -0.00001<br>(0.000004) | |
| constant | 0.46<br>(6.18) | |
| N | 240 | |
| Markets | 4 | |
| Halton draws | 100 | |
| f (p) ojective function value | 131.189 | |

[154] Table 3(a) shows the results of BLP estimation procedure with non-unified preference among borrowers. The estimation procedure strictly follows the simulation and GMM estimation procedure proposed by Berry et al. (1995). $\sum$ denotes the variance of the coefficient. The numbers in the parenthesis are the standard error of the estimators. This is a robustness check by using the random sample of small banks.

**Table 4(b) – t-test on interest rate elasticities by using a random sample**[155]

## *Non-parametric Comparison*

***Large banks v.s. Small banks***

| Group | Largest 30 | Smallest 30 (randomly selected) |
|---|---|---|
| Avg own-interest rate elasticity | 7.15 | 8.81 |
| std. error | 0.38 | 0.20 |
| N | 120 | 120 |
| Hypothesis | ***H0:*** own-interest rate elasticity ***(Largest 30) <*** own-interest rate elasticity ***(Smallest 30)*** | |
| t-stats | -3.85 | |
| Pr(T<t) | 0.0001 | |

[155] Table 4(b) shows the results of the t-test results between the largest 30 banks and the randomly selected 30 smallest banks. The null hypothesis is that the largest 30 banks are less elastic than the 30 smallest banks. The smallest banks are randomly selected from the 0-25% percentile in terms of the total value of assets.

## B.2 Data used for share computation

**Table 5 – Total book values of loans used for share computation**

| ***Total Book Values of outstanding Loans in the U.S.*** | |
|---|---|
| Year | Total of the industry[156] (Unit: Billion $) |
| 2013 | 84.52 |
| 2014 | 96.53 |
| 2015 | 116.37 |
| 2016 | 80.22 |

[156] Sources: FDIC quarterly reports.

# APPENDIX C. ROBUSTNESS CHECKS AND ADDITIONAL DATA OF CHAPTER THREE

## C.1 Robustness checks

### Table 1 – Robustness check with bootstrapping standard errors

**Bootstrapping for the main specifications**

| | *Logit 2007* | *Logit 2010* |
|---|---|---|
| **Variable** | **Coefficient** | **Coefficient** |
| ***LEVERAGE*** | **-0.8637*** | -0.01127 |
| | (0.4713) | (0.07687) |
| ***ASSET*** | -0.00108 | **-0.00372**** |
| | (0.00075) | (0.00156) |
| ***GROWTH*** | **0.526**** | **0.80932***** |
| | (0.2178) | (0.2528) |
| ***MARGIN*** | **23.727***** | -3.212 |
| | (6.98) | (3.753) |
| ***Tax/Rev*** | **-169.5436***** | 0.842 |
| | (58.861623) | (22.80) |
| ***CAPEX*** | 0.0001587 | **0.00481**** |
| | (0.0001024) | (0.00231) |
| ***D2*** | -1.2855 | -0.13391 |
| | (0.56992) | (0.51832) |
| ***D1*** | 1.3567 | 0.00782 |
| | 0.5630) | (0.6098) |
| ***INET*ICTAX*** | **1.356723**** | 0.01087 |
| | (0.5669) | (0.22488) |
| ***Equity*GOV*** | 0.0003501 | -0.00251 |
| | (0.00024) | (0.000736) |
| ***constant*** | **-4.856221**** | **-2.5528***** |
| | (2.218) | (0.9490) |
| ***log likelihood*** | -18.513 | -71.8165 |
| ***N*** | 164 | 409 |
| ***Pseudo R2*** | 0.8059 | 0.5601 |

## Table 2 – Robustness check with subsamples

**Subsample regression**

| | *Logit 2007* | *Logit 2010* |
|---|---|---|
| **Variable** | **Coefficient** | **Coefficient** |
| ***LEVERAGE*** | -0.5056 | 0.15488 |
| | (0.7257) | (0.12343) |
| ***ASSET*** | -0.00062 | **-0.0028***** |
| | (0.001268) | (0.0007315) |
| ***GROWTH*** | **0.507*** | **0.88841***** |
| | (0.2620) | (0.29512) |
| ***MARGIN*** | **21.21***** | **-7.08123*** |
| | (5.84) | (3.854) |
| ***Tax/Rev*** | **-143.57***** | 21.3684 |
| | (39.811) | (18.2762) |
| ***CAPEX*** | 0.0001045 | **0.003756**** |
| | (0.000370) | (0.00159) |
| ***D2*** | -1.213 | -0.03391 |
| | (1.2947) | (0.6083) |
| ***D1*** | | |
| | | |
| ***INET*ICTAX*** | **1.1757***** | 0.00934 |
| | (0.29834) | (0.00867) |
| ***Equity*GOV*** | 0.0001994 | **0.00190***** |
| | (0.000411) | 0.000413 |
| ***constant*** | **-4.746**** | **-2.9775***** |
| | (1.5415) | (0.78664) |
| ***log likelihood*** | -17.3425 | -64.6921 |
| ***N*** | 127 | 366 |
| ***Pseudo R2*** | 0.7762 | 0.5389 |

## Table 3 – Robustness check with probit models

| Probit regression | | |
|---|---|---|
| | ***Probit 2007*** | ***Probit 2010*** |
| **Variable** | **Coefficient** | **Coefficient** |
| ***LEVERAGE*** | -0.3832 | -0.0420 |
| | (0.32012) | (0.03932) |
| ***ASSET*** | -0.00065 | **-0.001513** |
| | (0.00062) | (0.0002513) |
| ***GROWTH*** | **0.2212*** | **0.3904***** |
| | (0.12887) | (0.1468) |
| ***MARGIN*** | **11.4509***** | -1.56784 |
| | (2.625) | (1.1623) |
| ***Tax/Rev*** | **-74.5275***** | -1.4058 |
| | (15.809) | (6.8803) |
| ***CAPEX*** | 0.0000542 | **0.001703***** |
| | (0.000169) | (0.00053) |
| ***D2*** | -0.90258 | 0.01621 |
| | (0.5568) | (0.28288) |
| ***D1*** | 0.55185 | 0.01576 |
| | (0.10045) | (0.39015) |
| ***INET*ICTAX*** | **0.55185***** | 0.005088 |
| | (0.100344) | (0.00444) |
| ***Equity*GOV*** | 0.0002063 | **0.0010492***** |
| | (0.0002014) | 0.000147 |
| ***constant*** | **-1.9038***** | **-1.4354***** |
| | (0.60798) | (0.33032) |
| ***log likelihood*** | -22.245 | -76.2515 |
| ***N*** | 164 | 409 |
| ***Pseudo R2*** | 0.7668 | 0.5331 |

**Table 4 – Robustness check on retained earnings**

**Logit regression with Retained Earnings**

| | *Logit 2007* | | *Logit 2010* |
|---|---|---|---|
| **Variable** | **Coefficient** | | **Coefficient** |
| ***LEVERAGE*** | -1.0361 | | -0.1173 |
| | (0.74951) | | (0.07206) |
| ***ASSET*** | -0.00133 | | **-0.00385*****|
| | (0.00131) | | (0.00071) |
| ***GROWTH*** | **0.4955**** | | **0.85042***** |
| | (0.2503) | | (0.25724) |
| ***MARGIN*** | **25.026***** | | -3.3894 |
| | (6.2821) | | (2.82892) |
| ***Tax/Rev*** | **-169.4795***** | | -2.6543 |
| | (40.4389) | | (15.8192) |
| ***CAPEX*** | 0.000202 | | **0.00526***** |
| | (0.000381) | | (0.00165) |
| ***D2*** | -1.28212 | | -0.09123 |
| | (1.246) | | (0.5902) |
| ***D1*** | -0.35275 | | 0.13039 |
| | (1.3245) | | (0.84955) |
| ***Retained Earnings*** | -0.00127 | | 0.00118 |
| | (0.001553) | | (0.000760) |
| ***INET*ICTAX*** | **1.3604***** | | 0.00966 |
| | (0.307) | | (0.00836) |
| ***Equity*GOV*** | 0.00044 | | **0.002526***** |
| | (0.000445) | | 0.0004006 |
| ***constant*** | **-4.6922***** | | **-2.62781***** |
| | (1.5360) | | (0.71165) |
| ***log likelihood*** | | -18.167 | -76.2515 |
| ***N*** | | 164 | 409 |
| ***Pseudo R2*** | | 0.8095 | 0.5331 |

## Table 5 – Robustness check with different proxies for government quality

**Logit regression with different proxies for government quality**

| | *(1) Stability* | *(2) Regulatory quality* | *(3) Corruption control* |
|---|---|---|---|
| **Variable** | **Coefficient** | **Coefficient** | **Coefficient** |
| ***LEVERAGE*** | -0.03245 | -0.0063 | -0.04573 |
| | (0.0715) | (0.0750) | (0.0691) |
| ***ASSET*** | **-0.00187*** | **-0.00142*** | **-0.00217*** |
| | (0.00052) | (0.000454) | (0.00054) |
| ***GROWTH*** | **0.605882** | **0.5335** | **0.6280*** |
| | (0.2503) | (0.2369) | (0.2400) |
| ***MARGIN*** | **-5.09401*** | **-5.4283** | **-4.8003*** |
| | (2.6078) | (2.4531) | (2.5573) |
| ***Tax/Rev*** | -13.22 | 15.52544 | 11.7703 |
| | (13.576) | (13.31607) | (13.758) |
| ***CAPEX*** | **0.002496** | **0.001941** | **0.0031** |
| | (0.001264) | (0.00072) | (0.00130) |
| ***D2*** | 0.208751 | 0.27595 | 0.15953 |
| | (0.52415) | (0.51682) | (0.53101) |
| ***D1*** | 0.47203 | 0.60730 | 0.3899 |
| | (0.59541) | (0.56262) | (0.6238) |
| ***INET*ICTAX*** | 0.002763 | 0.00342 | 0.00362 |
| | (0.00696) | (0.00675) | (0.00796) |
| ***Equity*GOV*** | **0.0014315*** | **0.0006253*** | **0.000913*** |
| | (0.001432) | (0.0001367) | 0.0001744 |
| ***constant*** | **-2.5801** | **-2.6481*** | **-2.5800*** |
| | (0.62778) | (0.6204) | (0.6335) |
| ***log likelihood*** | -100.3002 | -105.25883 | -96.0275 |
| ***N*** | 409 | 409 | 409 |
| ***Pseudo R2*** | 0.3859 | 0.3555 | 0.412 |

## C.2 Robustness checks

**Table 5 – Ex-post Analysis Data Summary Statistics and DID estimation**

| Ex-post Analysis Data Summary Statistics | | | | | |
|---|---|---|---|---|---|
| ***Variable*** | Obs | Mean | Std. Dev. | Min | Max |
| ***growth*** | 31,319 | 31.11% | 42.95% | -1434.98% | 1681.20% |
| ***Tax/revenue*** | 32,907 | 0.006028 | 0.811082 | -0.89 | 0.992 |
| ***asset*** | 36,916 | 1.77E+10 | 3.11E+11 | 0 | 1.75E+13 |
| ***CAPEX*** | 28,510 | 0.513799 | 0.684252 | -0.004515 | 0.8765 |
| ***ROA*** | 36,388 | 5.55% | 21.87% | -45.61% | 679.21% |
| ***ROE*** | 36,169 | -0.11% | 50.12% | -100.67% | 751.68% |
| ***Profit margin*** | 35,973 | 24.55% | 14.56% | -8.26% | 49.68% |
| ***Current ratio*** | 36,108 | 65.50% | 21.74% | 346.12% | 12.14% |
| ***Operational income to profit*** | 31,985 | 73.30% | 17.32% | 42.87% | 99.62% |

## C.3 Map of coastal provinces

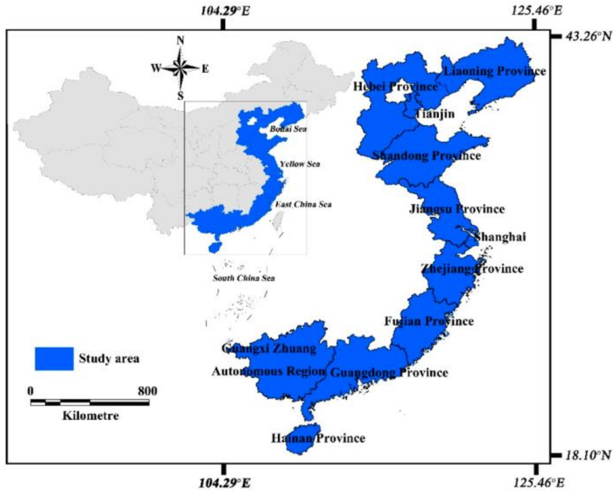


**Figure 1 – The provinces defined as the coastal area of China are in blue. Source: https://en.wikipedia.org/wiki/Coastline_of_China, last visit: 01/02/2017**

# REFERENCES


[1] Ackerberg, Daniel A., and Gautam Gowrisankaran. “Quantifying Equilibrium Network Externalities in the ACH Banking Industry.” The RAND Journal of Economics, vol. 37, no. 3, 2006, pp. 738–761.

[2] Addison, John T. "The Determinants of Firm Performance: Unions, Works Councils, and Employee Involvement/High-Performance Work Practices." Scottish Journal of Political Economy 52.3 (2005): 406-450.

[3] Alexander, Gordon J., Cheol S. Eun, and Sundaram Janakiramanan. "International listings and stock returns: Some empirical evidence." Journal of Financial and Quantitative Analysis 23.02 (1988): 135-151.

[4] Alexander, Gordon J., Cheol S. Eun, and Sundaram Janakiramanan. "Asset pricing and dual listing on foreign capital markets: A note." The Journal of finance 42.1 (1987): 151-158.

[5] Allison, Paul D. Missing data. Vol. 136. Sage publications, 2001.

[6] Almeida, H., M. Campello, B. Laranjeira, and S. Weisbenner, 2011, Corporate debt maturity and the real effects of the 2007 credit crisis, Critical Finance Review.

[7] Almossawi, Mohammed. "Bank selection criteria employed by college students in Bahrain: an empirical analysis." International Journal of Bank Marketing 19.3 (2001): 115-125.

[8] Ammendola, Giuseppe. "Devil Take the Hindmost: A History of Financial Speculation." American Foreign Policy Interests 22.1 (2000)

[9] Andersen, Erling Bernhard. "Asymptotic properties of conditional maximum-likelihood estimators." Journal of the Royal Statistical Society: Series B (Methodological) 32.2 (1970): 283-301.

[10] Anderson, Eugene W., Claes Fornell, and Sanal K. Mazvancheryl. "Customer satisfaction and shareholder value." Journal of marketing 68.4 (2004): 172-185.

[11] Anguo, Liu, Gao Ge, and Yang Kaizhong. "Returns to scale in the production of selected manufacturing sectors in China." Energy Procedia 5 (2011): 604-612

[12] Arellano, Manuel, and Stephen Bond. "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations." The review of economic studies 58.2 (1991): 277-297.

[13] Aslan, Hadiye and Kumar, Praveen, Import Competition and the Decline in U.S. Entrepreneurship (October 21, 2017). Available at SSRN: https://ssrn.com/abstract=3056827

[14] Aumann, Robert J., and Sergiu Hart, eds. Handbook of game theory with economic applications. Vol. 2. Elsevier, 1992.

[15] Autor, David, David Dorn, and Gordon H. Hanson. "The China syndrome: Local labor market effects of import competition in the US." American Economic Review 103.6 (2013): 2121-68.

[16] Autor, David H. "Structural demand shifts and potential labor supply responses in the new century." Federal Reserve Bank of Boston Conference Series. No. 52. 2007.

[17] Aydin, Dursun. "A comparison of the nonparametric regression models using smoothing spline and kernel regression." World Academy of Science, Engineering and Technology 36 (2007): 253-257.

[18] Baker, H. Kent, John R. Nofsinger, and Daniel G. Weaver. "International cross-listing and visibility." Journal of Financial and Quantitative Analysis37.03 (2002): 495-521.

[19] Barro, Robert J. "Economic Growth in a Cross Section of Countries." Quarterly Journal of Economics, May 1991, 106(2), pp. 407-43

[20] Becker, Gary S. "Human capital revisited." Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education (3rd Edition). The University of Chicago Press, 1994. 15-28.

[21] Berger, Allen N., et al. "The transformation of the US banking industry: What a long, strange trip it's been." Brookings papers on economic activity 1995.2 (1995): 55-218.

[22] Berger, Allen N., et al. "Bank concentration and competition: An evolution in the making." Journal of Money, Credit and Banking (2004): 433-451.

[23] Berkowitz, Daniel, Johannes Moenius, and Katharina Pistor. "Trade, law, and product complexity." The Review of Economics and Statistics 88.2 (2006): 363-373.

[24] Bernile, Gennaro, Alok Kumar, and Johan Sulaeman. "Home away from home: Geography of information and local investors." Review of Financial Studies (2015): hhv004.

[25] Berry, Steven, and Ariel Pakes. "The pure characteristics demand model." International Economic Review 48.4 (2007): 1193-1225.

[26] Berry, Steven, James Levinsohn, and Ariel Pakes. "Automobile prices in market equilibrium." Econometrica: Journal of the Econometric Society (1995): 841-890.

[27] Berry, Steven T. "Estimating discrete-choice models of product differentiation." The RAND Journal of Economics (1994): 242-262.

[28] Bhaumik, Sumon K., et al. "Does institutional quality affect firm performance? Insights from a semi-parametric approach." (2012).

[29] Bierman, L., & Gely, R. 1995. Striker replacements: A law, economics, and negotiations approach. South- ern California Law Review, 68: 363-396.

[30] Blass, Asher, and Yishay Yafeh. "Vagabond shoes longing to stray: Why foreign firms list in the United States." Journal of Banking & Finance 25.3 (2001): 555-572.

[31] Blundell, Richard, and James L. Powell. "Endogeneity in nonparametric and semiparametric regression models." Econometric society monographs 36 (2003): 312-357.

[32] Bontis, N. & Fitzenz, J. (2002). Intellectual capital ROI: A current map to human capital antecedents and consequences. Journal of Intellectual Capital, 3(3), 223-247.

[33] Boyd, William L., Myron Leonard, and Charles White. "Customer preferences for financial services: an analysis." International Journal of Bank Marketing 12.1 (1994): 9-15.

[34] Bresnahan, Timothy F. "Empirical studies of industries with market power." Handbook of industrial organization 2 (1989): 1011-1057.

[35] Bresnahan, Timothy F., and Peter C. Reiss. "Entry and competition in concentrated markets." Journal of Political Economy 99.5 (1991): 977-1009.

[36] Bresnahan, Timothy F., Scott Stern, and Manuel Trajtenberg. Market segmentation and the sources of rents from innovation: Personal computers in the late 1980's. No. w5726. National bureau of economic research, 1996.

[37] Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt. "Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence." The Quarterly Journal of Economics 117.1 (2002): 339-376.

[38] Cetorelli, Nicola, and Philip E. Strahan. "Finance as a barrier to entry: Bank competition and industry structure in local US markets." The Journal of Finance 61.1 (2006): 437-461.

[39] Chan, Kalok, Albert J. Menkveld, and Zhishu Yang. "Information asymmetry and asset prices: Evidence from the China foreign share discount." The Journal of Finance 63.1 (2008): 159-196.

[40] Chandler, Gaylen N., and Douglas W. Lyon. "Involvement in knowledge-acquisition activities by venture team members and venture performance." Entrepreneurship Theory and Practice33.3 (2009): 571-592.

[41] Chen, David Yunchao. "China’s mass higher education: Problem, analysis, and solutions." Asia Pacific Education Review5.1 (2004): 23.

[42] Chevalier, Judith A. “Capital Structure and Product-Market Competition: Empirical Evidence from the Supermarket Industry.” The American Economic Review, vol. 85, no. 3, 1995, pp. 415–435.

[43] Chi, Jing, and Carol Padgett. "THE PERFORMANCE AND LONG-RUN CHARACTERISTICS OF THE CHINESE IPO MARKET." Pacific Economic Review 10.4 (2005): 451-469. applications. Cambridge, MA: Mit Press, 1981.

[44] Choi, Jong-Hag, et al. "Cross-listing audit fee premiums: Theory and evidence." The Accounting Review 84.5 (2009): 1429-1463.

[45] Churchill, Sefa Awaworyi, Maria Rebecca Valenzuela, and Wisdom Sablah. "Ethnic diversity and firm performance: Evidence from China's materials and industrial sectors." Empirical Economics 53.4 (2017): 1711-1731.

[46] Crook, T. Russell, et al. "Does human capital matter? A meta-analysis of the relationship between human capital and firm performance." Journal of applied psychology 96.3 (2011): 443.

[47] Dang, Chongyu, Zhichuan Frank Li, and Chen Yang. "Measuring firm size in empirical corporate finance." Journal of Banking & Finance 86 (2018): 159-176.

[48] Daniel Kaufmann, Aart Kraay and Massimo Mastruzzi (2010). "The Worldwide Governance Indicators : A Summary of Methodology, Data and Analytical Issues". World Bank Policy Research Working Paper No. 5430

[49] D'Aveni, R. A. 1989. The aftermath of organizational decline: A longitudinal study of the strategic and managerial characteristics of declining firms. Acad- emy of Management Journal, 32: 577-605.

[50] Davis, Peter. "Spatial competition in retail markets: movie theaters." The RAND Journal of Economics 37.4 (2006): 964-982.

[51] De Boor, Carl, et al. A practical guide to splines. Vol. 27. New York: Springer-Verlag, 1978.

[52] De Carolis, Donna Marie, Barrie E. Litzky, and Kimberly A. Eddleston. "Why networks enhance the progress of new venture creation: The influence of social capital and cognition." Entrepreneurship theory and practice 33.2 (2009): 527-545.

[53] Doidge, Craig, G. Andrew Karolyi, and René M. Stulz. "Why are foreign firms listed in the US worth more?." Journal of financial economics 71.2 (2004): 205-238.

[54] Doidge, Craig, et al. "Private benefits of control, ownership, and the cross-listing decision." The Journal of Finance 64.1 (2009): 425-466.

[55] Doong, Shuh-Chyi, Hung-Gay Fung, and Jr-Ya Wu. "Are social, financial, and human capital value enhancing? Evidence from Taiwanese firms." International Review of Economics & Finance20.3 (2011): 395-405.

[56] Fasci, Martha A., and Jude Valdez. "A performance contrast of male-and female-owned small accounting practices." Journal of small business management 36.3 (1998): 1.

[57] Fernald, John, and John H. Rogers. “Puzzles in the Chinese Stock Market.” The Review of Economics and Statistics, vol. 84, no. 3, 2002, pp. 416–432. JSTOR, JSTOR,

[58] Fleisher, Belton, Haizheng Li, and Min Qiang Zhao. "Human capital, economic growth, and regional inequality in China." Journal of development economics 92.2 (2010): 215-231.

[59] Foster, Andrew D., and Mark R. Rosenzweig. "Learning by doing and learning from others: Human capital and technical change in agriculture." Journal of political Economy 103.6 (1995): 1176-1209.

[60] Fraser, Donald R., Jeff Madura, and Robert A. Weigand. "Sources of bank interest rate risk." Financial Review 37.3 (2002): 351-367.

[61] Friedman, Jerome H. "Multivariate adaptive regression splines." The annals of statistics 19.1 (1991): 1-67.

[62] Gaud, Philippe, Martin Hoesli, and André Bender. "Debt-equity choice in Europe." International Review of Financial Analysis 16.3 (2007): 201-222.

[63] Ghose, A., P. G. Ipeirotis, B. Li. 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. Marketing Science 31 (3) 493-520.

[64] Gilbert, R. Alton. "Bank market structure and competition: a survey." Journal of Money, Credit and Banking 16.4 (1984): 617-645.

[65] Giroud, Xavier, and Holger M. Mueller. "Does corporate governance matter in competitive industries?." Journal of Financial Economics 95.3 (2010): 312-331.

[66] Goodfriend, Marvin. "Interest rates and the conduct of monetary policy." Carnegie-Rochester conference series on public policy. Vol. 34. North-Holland, 1991.

[67] Grant, R. M. 1996. Toward a knowledge-based theory of the firm. Strategic Management Journal, 1 7(special issue): 109-122.

[68] Hal, R. Varian. Microeconomic analysis. 2016.

[69] Hamermesh, Daniel S., and James Grant. "Econometric studies of labor-labor substitution and their implications for policy." Journal of Human Resources (1979): 518-542.

[70] Hannan, Timothy H. "Market share inequality, the number of competitors, and the HHI: An examination of bank pricing." Review of Industrial Organization 12.1 (1997): 23-35.

[71] Hansen, Gary S., and Birger Wernerfelt. "Determinants of firm performance: The relative importance of economic and organizational factors." Strategic management journal 10.5 (1989): 399-411.

[72] Hanushek, Eric A., and Dennis D. Kimko. "Schooling, labor-force quality, and the growth of nations." American economic review (2000): 1184-1208.

[73] Hausman, Jerry A. "Specification tests in econometrics." Econometrica: Journal of the Econometric Society (1978): 1251-1271.

[74] Hau, Harald. "The role of transaction costs for financial volatility: Evidence from the Paris Bourse." Journal of the European Economic Association 4.4 (2006): 862-890.

[75] Hausman, Jerry, and Daniel McFadden. "Specification tests for the multinomial logit model." Econometrica: Journal of the Econometric Society (1984): 1219-1240.

[76] Hausman, Jerry, Gregory Leonard, and J. Douglas Zona. "Competitive analysis with differenciated products." Annales d'Economie et de Statistique (1994): 159-180.

[77] Hausman, Jerry A., and Gregory K. Leonard. "The competitive effects of a new product introduction: A case study." The Journal of Industrial Economics 50.3 (2002): 237-263.

[78] Hausman, Jerry A., and William E. Taylor. "Panel data and unobservable individual effects." Econometrica: Journal of the Econometric Society (1981): 1377-1398.

[79] Hausman, Jerry, Gregory Leonard, and J. Douglas Zona. "Competitive analysis with differenciated products." Annales d'Economie et de Statistique (1994): 159-180.

[80] Heckman, James J. "Policies to foster human capital." Research in economics 54.1 (2000): 3-56.

[81] Hillman, Amy J., and Thomas Dalziel. "Boards of directors and firm performance: Integrating agency and resource dependence perspectives." Academy of Management review 28.3 (2003): 383-396.

[82] Hitt, Michael A., et al. "Direct and moderating effects of human capital on strategy and performance in professional service firms: A resource-based perspective." Academy of Management journal 44.1 (2001): 13-28.

[83] Ho, Chun-Yu. "Switching cost and deposit demand in China." International Economic Review 56.3 (2015): 723-749.

[84] Ittner, Christopher D., and David F. Larcker. "Are nonfinancial measures leading indicators of financial performance? An analysis of customer satisfaction." Journal of accounting research 36 (1998): 1-35.

[85] Jones, Benjamin F. "The human capital stock: a generalized approach." American Economic Review 104.11 (2014): 3752-77.

[86] Judson, Ruth A., and Ann L. Owen. "Estimating dynamic panel data models: a guide for macroeconomists." Economics letters65.1 (1999): 9-15.

[87] Kaplan, Steven N., and Luigi Zingales. "Do investment-cash flow sensitivities provide useful measures of financing constraints?." The Quarterly Journal of Economics 112.1 (1997): 169-215.

[88] Kaufmann, George G. "The diminishing role of commercial banking in the US economy." The crisis in American banking (1993): 139-159.

[89] Kaynak, Erdener, and Orsay Kucukemiroglu. "Bank and product selection: Hong Kong." International Journal of Bank Marketing 10.1 (1992): 3-16.

[90] Keeley, Michael C. "Deposit insurance, risk, and market power in banking." The American Economic Review (1990): 1183-1200.

[91] Kennington, Carolyn, Jeanne Hill, and Anna Rakowska. "Consumer selection criteria for banks in Poland." International Journal of Bank Marketing 14.4 (1996): 12-21.

[92] Kim, Moshe, Doron Kliger, and Bent Vale. "Estimating switching costs: the case of banking." Journal of Financial Intermediation 12.1 (2003): 25-56.

[93] Kluve, Jochen, Hartmut Lehmann, and Christoph M. Schmidt. "Active labor market policies in Poland: human capital enhancement, stigmatization, or benefit churning?." Journal of Comparative Economics 27.1 (1999): 61-89.

[94] Kor, Yasemin Y., and Huseyin Leblebici. "How do interdependencies among human-capital deployment, development, and diversification strategies affect firms' financial performance?." Strategic Management Journal 26.10 (2005): 967-985.

[95] Krueger, Alan B., and Mikael Lindahl. Education for growth: Why and for whom?. No. w7591. National Bureau of Economic Research, 2000.

[96] Lane, P. J., & Lubatkin, M. 1998. Relative absorptive capacity and interorganizational learning. Strategic ManagementJournal, 19: 461-477.

[97] Lang, Mark H., Karl V. Lins, and Darius P. Miller. "ADRs, analysts, and accuracy: Does cross listing in the United States improve a firm's information environment and increase market value?." Journal of Accounting Research 41.2 (2003): 317-345.

[98] La Porta, Rafael, et al. "The quality of government." Journal of Law, Economics, and organization 15.1 (1999): 222-279.

[99] Leonard, Jonathan S. "Executive pay and firm performance." ILR Review 43.3 (1990): 13-S.

[100] Levchenko, Andrei A. "Institutional quality and international trade." The Review of Economic Studies 74.3 (2007): 791-819.

[101] Levine, Ross and Renelt, David. "A Sensitivity Analysis of Cross-Country Growth Regres- sions." American Economic Review, Septem- ber 1992, 82(4), pp. 942-63.

[102] Levine, Ross and Zervos, Sara J. "What We Have Learned About Policy and Growth From Cross-Country Regressions?" American Economic Review, May 1993 (Papers and Pro- ceedings), 83(2), pp. 426-30

[103] Li, Hongbin, et al. "Political connections, financing and firm performance: Evidence from Chinese private firms." Journal of development economics87.2 (2008): 283-299.

[104] Licht, Amir N. "Cross-listing and corporate governance: bonding or avoiding." Chi. J. Int'l L. 4 (2003): 141.

[105] Liebeskind, J. P. 1996. Knowledge, strategy and the theory of the firm. Strategic Management Journal, 17(special issue): 93-107.

[106] MacKay, Peter, and Gordon M. Phillips. "How does industry affect firm financial structure?." The Review of Financial Studies 18.4 (2005): 1433-1466.

[107] Maksimovic, Vojislav, and Josef Zechner. "Debt, agency costs, and industry equilibrium." The Journal of Finance 46.5 (1991): 1619-1643.

[108] Mankiw, N. Gregory; Romer, David and Weil, David. "A Contribution to the Empirics of Economic Growth." Quarterly Journal of Economics, May 1992, 107(2), pp. 407-37.

[109] Manski, Charles F., and Daniel McFadden. "Alternative estimators and sample designs for discrete choice analysis." Structural analysis of discrete data with econometric applications 2 (1981).

[110] McFadden, Daniel. "Conditional logit analysis of qualitative choice behavior." (1973): 105.

[111] McFadden, Daniel. "Econometric models for probabilistic choice among products." Journal of Business (1980): S13-S29.

[112] Merton, Robert C. "A simple model of capital market equilibrium with incomplete information." The journal of finance 42.3 (1987): 483-510.[33

[113] Miller, Danny, and Le Breton-Miller. "Family governance and firm performance: Agency, stewardship, and capabilities." Family business review19.1 (2006): 73-87.

[114] Morck, Randall, Bernard Yeung, and Wayne Yu. "The information content of stock markets: why do emerging markets have synchronous stock price movements?." Journal of financial economics 58.1 (2000): 215-260.

[115] Murphy, Kevin J. "Corporate performance and managerial remuneration: An empirical analysis." Journal of accounting and economics 7.1-3 (1985): 11-42.

[116] Murray, Alan I. "Top management group heterogeneity and firm performance." Strategic management journal 10.S1 (1989): 125-141.

[117] Neely, Andy, ed. Business performance measurement: theory and practice. Cambridge University Press, 2002.

[118] Nelson, Richard R. and Phelps, Edmund. "Invest-ment in Humans, Technology Diffusion, and Economic Growth." American Economic Re-view, May 1966 (Papers and Proceedings), 56(2), pp. 69-75.

[119] Nevo, Aviv. "Mergers with differentiated products: The case of the ready-to-eat cereal industry." The RAND Journal of Economics (2000): 395-421.

[120] Ng, Thomas WH, and Daniel C. Feldman. "Human capital and objective indicators of career success: The mediating effects of cognitive ability and conscientiousness." Journal of Occupational and Organizational Psychology 83.1 (2010): 207-235.

[121] Pagano, Marco, Fabio Panetta, and Luigi Zingales. "Why do companies go public? An empirical analysis." The Journal of Finance 53.1 (1998): 27-64.

[122] Pagano, Marco, et al. "What makes stock exchanges succeed? Evidence from cross-listing decisions." European Economic Review 45.4 (2001): 770-782.

[123] Pascual, Roberto, Bartolomé Pascual-Fuster, and Francisco Climent. "Cross-listing, price discovery and the informativeness of the trading process." Journal of Financial Markets 9.2 (2006): 144-161.

[124] Paul Krugman, Robin Wells Macroeconomics 4th edition (2015), Macmillan Publishers. ISBN-10: 1429283432

[125] Pennings, Johannes M., et al. “Human Capital, Social Capital, and Firm Dissolution.” The Academy of Management Journal, vol. 41, no. 4, 1998, pp. 425–440.

[126] Petersen, Mitchell A. "Estimating standard errors in finance panel data sets: Comparing approaches." The Review of Financial Studies 22.1 (2009): 435-480.

[127] Petrin, Amil, and Kenneth Train. "A control function approach to endogeneity in consumer choice models." Journal of marketing research 47.1 (2010): 3-13.

[128] Pinheiro, J. C., and D. M. Bates. 2000. Mixed-Effects Models in S and S-PLUS. New York: Springer

[129] Polanyi, M. 1967. The tacit dimension. Garden City, NY: Anchor Publishing.

[130] Powell, James L. "Endogeneity and Instruments in Nonparametric Models." Advances in Economics and Econometrics (2003): 358.

[131] Phillips, Gordon M. "Increased debt and industry product markets an empirical analysis." Journal of financial Economics37.2 (1995): 189-238.

[132] Phuong Ta, Huu, and Kar Yin Har. "A study of bank selection decisions in Singapore using the analytical hierarchy process." International Journal of Bank Marketing 18.4 (2000): 170-180.

[133] Pástor, Ľuboš, and Robert F. Stambaugh. "Comparing asset pricing models: an investment perspective." Journal of Financial Economics 56.3 (2000): 335-381.

[134] Rajan, Raghuram G. "Insiders and outsiders: The choice between informed and arm's-length debt." The Journal of finance 47.4 (1992): 1367-1400.

[135] Rebelo, Sergio. "Long-Run Policy Analysis and Long-Run Growth." Journal of Political Economy, June 1991, 99(3), pp. 500-21.

[136] Riahi-Belkaoui, Ahmed. "Intellectual capital and firm performance of US multinational firms: a study of the resource-based and stakeholder views." Journal of Intellectual capital 4.2 (2003): 215-226.

[137] Romer, Paul. "Endogenous Technological Change." Journal of Political Economy, October 1990a, Pt. 2, 99(5), pp. S71-S102. . "Human Capital and Growth: Theory and Evidence." Carnegie-Rochester Confer-ence Series on Public Policy, Spring 1990b, 32, pp. 251-86.

[138] Salas, Vicente, and Jesus Saurina. "Deregulation, market power and risk behaviour in Spanish banks." European Economic Review 47.6 (2003): 1061-1075.

[139] Sami, Heibatollah, and Haiyan Zhou. "The Economic Consequences of Increased Disclosure: Evidence from Cross-Listings of Chinese Firms." Journal of International Financial Management & Accounting 19.1 (2008): 1-27.

[140] Sarkissian, Sergei, and Michael J. Schill. "The overseas listing decision: New evidence of proximity preference." Review of Financial Studies 17.3 (2004): 769-809.

[141] Schultz, Theodore W. "Investment in Human Capital. The Role of Education and of Research." (1971).

[142] Shaffer, Sherrill. "Competition in the US banking industry." Economics letters 29.4 (1989): 321-323.

[143] Shaw, Kathryn L. "Life-cycle labor supply with human capital accumulation." International Economic Review (1989): 431-456.

[144] Shrader, Rod, and Donald S. Siegel. "Assessing the relationship between human capital and firm performance: evidence from technology-based new ventures." Entrepreneurship Theory and Practice 31.6 (2007): 893-908.

[145] Shy, Oz. "A quick-and-easy method for estimating switching costs." International Journal of Industrial Organization 20.1 (2002): 71-87.

[146] Smart, John C. "Gender equity in academic rank and salary." The review of higher education 14.4 (1991): 511-525.

[147] Stavins, Robert N. "Transaction costs and tradeable permits." Journal of environmental economics and management 29.2 (1995): 133-148.

[148] Stephens, Matthew, and Paul Scheet. "Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation." The American Journal of Human Genetics 76.3 (2005): 449-462.

[149] Stierwald, Andreas. "Determinants of firm profitability-the effect of productivity and its persistence." Melbourne Institute of Applied Economic and Social Research (2009): 1-25.

[150] Song, Minjae. "A hybrid discrete choice model of differentiated product demand with an application to personal computers." International Economic Review 56.1 (2015): 265-301.

[151] Sudhir, Karunakaran. "Structural analysis of manufacturer pricing in the presence of a strategic retailer." Marketing Science 20.3 (2001): 244-264

[152] Tibshirani, Robert (1997). "The lasso Method for Variable Selection in the Cox Model". Statistics in Medicine. 16: 385–395

[153] Tirole, Jean, "The Theory of Industrial Organization", Cambridge, Massachusetts: The MIT Press, 1988

[154] Ton, Zeynep. "The effect of labor on profitability: The role of quality." (2009).

[155] Train, Kenneth. Discrete choice methods with simulation. Cambridge university press, 2003.

[156] Varian, H; Microeconomic analysis. (3rd ed.)Norton, New York, NY (1992). Copyright © 1995 Published by Elsevier BV

[157] Vermunt, Jeroen K., and Jay Magidson. "Latent class cluster analysis." Applied latent class analysis 11 (2002): 89-106.

[158] Villalonga, Belén. "Intangible resources, Tobin'sq, and sustainability of performance differences." Journal of Economic Behavior & Organization 54.2 (2004): 205-230.

[159] Wagner, R. Harrison. "Economic interdependence, bargaining power, and political influence." International Organization 42.3 (1988): 461-483.

[160] Wang, Yan, and Yudong Yao. Sources of China's economic growth, 1952–99: incorporating human capital accumulation. The World Bank, 1999.

[161] Werner, Ingrid M., and Allan W. Kleidon. "UK and US trading of British cross-listed stocks: An intraday analysis of market integration." Review of Financial Studies 9.2 (1996): 619-664.

[162] Wooldridge, Jeffrey M. Econometric analysis of cross section and panel data. MIT press, 2010.

[163] Xu, Yuqian, Mor Armony, and Anindya Ghose. "The effect of online reviews on physician demand: A structural model of patient choice." (2016).

[164] Yang, Dennis Tao. "Education in production: measuring labor quality and management." American Journal of Agricultural Economics 79.3 (1997): 764-772.

[165] Zineldin, Mosad. "Bank strategic positioning and some determinants of bank selection." International Journal of Bank Marketing 14.6 (1996): 12-22.